НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут телекомунікаційних систем
Кафедра Інформаційно-телекомунікаційних мереж

«На правах рукопису»
УДК _____

«До захисту допущено»
Завідувач кафедри
_____ Лариса ГЛОБА
«___»_____2020 р.

Магістерська дисертація
на здобуття ступеня магістра
за освітньо-професійною програмою «Інформаційно-комунікаційні технології»
зі спеціальності 172 «Телекомунікації та радіотехніка»

на тему: «Модифікований метод машинного навчання для обробки даних в мережі MicroGrid»

Виконав:
студент VI курсу, групи ТІ-91мп
Блазнов Володимир Миколайович _____

Керівник:
Старший викладач кафедри ІТМ ІТС, к.т.н
Суліма Світлана Валеріївна _____

Консультант:
Асистент кафедри ІТМ ІТС
Курдеча Василь Васильович _____

Рецензент:
Зав. кафедри промислової електроніки
КПІ ім. Ігоря Сікорського, проф., д.т.н.
Ямненко Юлія Сергіївна _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.

Студент _____

Київ – 2020 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Інститут телекомунікаційних систем
Кафедра Інформаційно-телекомунікаційних мереж

Рівень вищої освіти – другий (магістерський)

Спеціальність – 172 «Телекомунікації та радіотехніка»

Освітньо-професійна програма «Інформаційно-комунікаційні технології»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Лариса ГЛОБА
«___»_____2020 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Блазнову Володимиру Миколайовичу

1. Тема дисертації «Модифікований метод машинного навчання для обробки даних в мережі MicroGrid», науковий керівник науковий керівник дисертації старший викладач кафедри інформаційно-телекомунікаційних мереж Суліма Світлана Валеріївна, к.т.н., затверджені наказом по університету від «03» листопада 2020 р. № 3208-с

2. Термін подання студентом дисертації 10.12.2020 р.

3. Об’єкт дослідження: Архітектура MicroGrid

4. Предмет дослідження: Обробка великих даних для прогнозування навантаження.

5. Перелік завдань, які потрібно розробити:

5.1 Аналіз існуючих рішень обробки даних.
5.2 Дослідження алгоритмів Machine Learning.
5.3 Метод прогнозування навантаження.
5.4 Опис системи обробки даних

6. Орієнтовний перелік ілюстративного матеріалу

7. Орієнтовний перелік публікацій

8. Консультанти розділів дисертації*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Курдеча В.В. асистент кафедри		
2	Курдеча В.В. асистент кафедри		
3	Курдеча В.В. асистент кафедри		
4	Курдеча В.В. асистент кафедри		
5	Курдеча В.В. асистент кафедри		

9. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Аналіз існуючих рішень обробки даних у MicroGrid	03.09.2020-20.09.2020	виконано
2	Дослідження існуючих алгоритмів у Machine Learning.	21.09.2020-10.10.2020	виконано
3	Прогнозування навантаження в системах MicroGrid	11.10.2020-24.10.2020	виконано
4	Опис методу обробки даних, вибір основних параметрів.	28.10.2020-28.11.2020	виконано
5	Розробка стартап проекту	29.11.2020 – 04.12.2020	виконано

Студент Володимир БЛАЗНОВ

Науковий керівник дисертації Світлана СУЛІМА

РЕФЕРАТ

Робота містить 76 сторіноки, 19 рисунків та 2 таблиці. Було використано 17 джерел.

Мета роботи. Метою даної роботи є підвищення ефективності обробки великих даних у MicroGrid за рахунок методу машинного навчання, що відрізняється гібридним методом прогнозування навантаження в системі.

В магістерскій десиртації розглядається задача застосування Anomaly Detection та Future Selection як гібридний засіб Machine Learning для обробки великих даних та прогнозування навантаження системи. Наукова новизна отриманих результатів полягає у отриманні подальшого розвитку теорії застосування методів машинного навчання для обробки та аналізу великих даних (Big Data) в частині прогнозування навантеження в системах MicroGrid[1]. Запропоновано розв’язання задачі прогнозування навантаження застосування методів опорних векторів (SVM) та Future Selection. Проведено аналіз та порівняння з іншими методами та виявлено переваги даного методу.

Ключові слова: MicroGrid, Big Data, Machine Learning, Anomaly Detection, Feature Selection.

ABSTRACT

The work contains 76 pages, 17 figures and 2 tables. 17 sources were used.

Goal. The purpose of this work is to increase the efficiency of big data processing in MicroGrid due to the method of machine learning, which differs from the hybrid method of load prediction in the system.

The master's dissertation considers the problem of using Anomaly Detection and Future Selection as a hybrid tool for Machine Learning to process large data and predict system load. The scientific novelty of the obtained results is to obtain further development of the theory of application of machine learning methods for processing and analysis of big data (Big Data) in terms of load prediction in MicroGrid systems [1]. The solution of the problem of load prediction using the methods of reference vectors (SVM) and Future Selection is proposed. The analysis and comparison with other methods are carried out and the advantages of this method are revealed.

Keywords: MicroGrid, Big Data, Machine Learning, Anomaly Detection, Feature Selection.

Зміст

ВСТУП

Актуальність. Розвиток електронних систем, інтелектуальних засобів керування та обміну інформацію призвів до формування нової інформаційно-енергетичної концепції SmartGrid, в рамках якої розглядаються локальні електротехнічні об’єкти – MicroGrid, що містять певний набір джерел та навантажень, які, як правило, підключені до централізованої електромережі. В рамках MicroGrid постає цілий ряд задач керування та узгодженого функціонування пристроїв, що призводить до необхідності оперування великими інформаційними потоками.

У галузі електроенергетики відбувається активне впровадження «інтелектуальних мереж» SmartGrid та MicroGrid. Тому тематика досліджень, присвячена обробці великих даних у MicroGrid та прогнозуванню навантеження за допомогою Machine Learning, є актуальною.

Мета роботи. Метою даної роботи є підвищення ефективності обробки великих даних у MicroGrid за рахунок гібридного методу машинного навчання, що відрізняється виявленням аномальної в системі та прогнозуванням навантеження на основі зібраних даних.

Для досягнення мети в роботі вирішуються такі задачі:

1. Аналіз існуючих рішень обробки даних у MicroGrid.
2. Дослідження існуючих алгоритмів у Machine Learning.
3. Опис методу прогнозування навантаження
4. Опис методу обробки даних, вибір основних параметрів.
5. Модифікований метод обробки даних.

Об’єктом дослідження є обробка великих даних, пов’язаних з аномальними даними які надходять з датчиків для покращення результатів прогнозування навантеження в системах MicroGrid.

Предметом дослідження є метод обробки даних за допомогою machine learning.

РОЗДІЛ 1

ОБРОБКА ВЕЛИКИХ ДАНИХ У СЕРЕДОВИЩІ MICROGRID

1.1. Існуючі рішення обробки даних в MicroGrid

Одним із відомих проектів розробки розумних будинків є "ACHE MicroGrid" в Боулдері, штат Колорадо, який використовує адаптивний контроль домашнього середовища та нейронних мереж[3].

ACHE MicroGrid контролює основні функції будинку, включаючи температуру, освітлення та опалення, без потреби втручання користувача. Метою ACHE MicroGrid є прогнозування та адаптація режимів роботи пристроїв будинку, розробка сервісу, заснованого на спостереженнях нейронної мережі за способос життя людей в будинку.

Проект Техаського університету MavHome[4], є розробкою яка обумовлює себе як "раціональний агент". Раціональний агент систематизує зібрані дані, обробляє їх, та намагається використати для вгадування дій користувача. Система MavHome базується на алгоритмі LeZi-update який відстежує користувачів та їх поведінкової моделі.

Проект "GatorTech MicroGrid" був розроблений у Флориді. Він складається з окремих пристроїв з можливістю обробки інформації, таких як ліжко, поштова скринька, підлога, вхідні двері та інші. Всі пристрої підключені до однієї платформи і маю датчики та виконавчі пристрої.

TERVA система особистого оздоровчого спостереження розроблена у Фінляндії дозволяє спостерігати за «змінами, пов'язаними з оздоровчою діяльністю вдома» в будь які проміжки часу. TERVA працює на домашньому обчислювальному пристрої та підтримує зв'язок з різними приладами контролюючими датчики в будинку. Система відстежує параметри артеріального тиску, температури користувача, ваги, інтервал серцебиття, частоту дихання, рухів тощо.

Microgrid була розроблена у Фінляндії, використовуючи внутрішнє відстеження та ємнісне позиціонування. Мета - виявити діяльність людини, яка живе в інтелектуальному будинку. Система використовує електроди, вбудовані в підлогу будинку, щоб виявити людину, а потім виявляє взаємодію людини з предметами побуту (столи, ліжка, холодильники та дивани).

Можна використовувати радіочастотні сигнали, а також датчики ліжка для виявлення присутності мешканців, яким доводиться вставати з ліжка у літньої людини.

Проект SPHERE здійснює моніторинг та поїзди всередині житлових будинків. Суть цього проекту полягає у додаванні даних датчиків та створенні набору даних для моніторингу та управління будь-яким станом здоров'я. SPHERE використовує фізичний (несучий) моніторинг сигналів, моніторинг домашнього середовища та моніторинг на основі філософії. Мультимодальний підхід до звукової системи повністю інтегрований з інтелектуальними алгоритмами обробки даних, які контролюють збір даних.

1.2. Джерела великих даних та їх реєстрація

Система MicroGrid відрізняється від повністю автоматизованої системи. Мікромережа характеризується присутністю людини, цей фактор призводить до необхідності врахування зовнішніх факторів - впливу людини на роботу обладнання та працездатність системи. Людський фактор призводить до роботи з великим обсягом даних. Тому розробники та дослідники змушені використовувати методи машинного навчання для обробки великих даних у своїй роботі. Велика різноманітність цих методів включає підходи, інструменти та принципи обробки, особливо ефективні в системах з постійним приростом даних, та необхідністю перетворення результатів, зрозумілих для сприйняття людиною. Методи машинного

навчання (Machine Learning) це важливий та ефективний інструмент обробки великих обсягів даних. Як галузь інформатики, Machine Learning використовує статистичні методи для забезпечення навчання комп'ютерних систем із поступовим покращенням продуктивності вирішення конкретних задач без явного програмування.

У MicroGrid розрізняють декілька аномальних режимів:

1. Аварійні випадки;

2. вихід технічних параметрів за допустимі межі;

3. Нетипова активність, незрозуміла поведінка людини – «людський фактор»

Перші дві справи вирішують суто технічні, інженерні та відділи інженерного розвитку. З точки зору машинного навчання в мікромережі, найбільший інтерес представляє третій випадок, оскільки наявність «людського фактора» призводить до великої кількості потенційних завдань, які слід обробляти як навчальну вибірку. Такі системи можуть містити великі обсяги даних, які неможливо обробити стандартними методами.

З багатьох методів машинного навчання найбільш підходящим є метод Anomaly Detection. У цій дисертації розглядається проблема використання виявлення аномалій у Microgrid як інструменту машинного навчання для обробки великих даних, зібраних під час спостереження та обробки даних.

Застосування технологій обробки великих даних, зокрема методів машинного навчання, за останні роки суттєво прискорило і поліпшило процес обробки даних практично в будь-якій предметній області (фізика, економіка, медицина і т.д.). Хоча в області управління електричними об'єктами та інтелектуальними енергетичними мережами, які зазвичай включають мікромережеві системи відновлюваних та альтернативних джерел енергії, дотепер такі методи не використовувались, хоча

формування алгоритмів управління з точки зору великих даних. Такі системи мають великий потенціал (погодні параметри, екологічні параметри, економічні показники). Нарешті, це дозволяє проводити детальний аналіз, аналіз та прогнозування режимів роботи, вибирати та проводити порівняльний аналіз найкращих стратегій управління, постійну діагностику станів та виявлення аномальних, передсхожих та надзвичайних ситуацій режимів роботи Це надзвичайно важливо Підвищення енергоефективності, всебічне впровадження та якість роботи MicroGrid. З іншого боку, враховуючи великий обсяг даних, потреба в подальшій обробці, зберіганні та передачі передбачає впровадження принципово нових методів та інформаційних технологій - хмарних та розподілених комп'ютерів, штучного інтелекту, машинного навчання та Інтернету речей. І таблиця єдиного інформаційного середовища.

Для сучасних скадних комплексів MicroGrid актуальною є проблема створення гетерогенної мережі збору, обробки та передавання даних, що характеризують внутрішній стан та параметри оточуючого середовища електротехнічних комплексів з джерелами розосередженої генерації MicroGrid: параметри клімату та екологічні параметри, технічні та технологічні параметри MicroGrid (фізичні та електричні параметри, показники ефективності режимів функціонування), економічні показники ефективності роботи (оптимізаційна цільова функція, динамічні тарифи на електричну енергію у MicroGrid). Враховуючи великий обсяг цих даних, що обґрунтовує віднесення їх до категорії BigData, можна стверджувати, що питання високоефективної та швидкодіючої обробки цих даних з використанням сучасних методів машинного навчання, розподілених обчислень, штучного інтелекту є надзвичайно актуальними. Такий підхід надає можливість забезпечити функціонування MicroGrid, зокрема, об'єктів військового призначення, критичними вимогами для яких є

мінімальна кількість фізичних інформаційних ліній, надійне електропостачання та можливість роботи в автономному режимі.

Сучасні електротехнічні комплекси із джерелами розосередженої генерації MicroGrid характеризуються високим ступенем насиченості різноманітними пристроями генерації, споживання та накопичення енергії, а також передавання великих масивів різнотипних (гетерогенних) даних, що значно ускладнює задачу побудови адекватних методів оперативного керування та інформаційного обміну у інформаційно-керуючому середовищі таких комплексів.

Технології обробки великих даних – методи машинного навчання, класифікації, аномальної детекції, регресійного аналізу – дозволяють оперувати з великими обсягами даних, накопичених в процесі тривалого функціонування MicroGrid в різних режимах, та є основою для вибору ефективної стратегії керування.

Для контролю за діяльністю мешканця в будинку використовується низка сенсорних пристроїв. Інформація, зібрана з пристроїв, обробляється та зберігається для аналізу та використання в поточному та майбутньому стані. Оскільки обсяг цих даних надзвичайно великим, це пояснює доцільність використання методів машинного навчання для обробки Big Data.

MicroGrid з вбудованим модулем інтелектуального машиннгоо навчання реалізує концепцію, в якій мережа датчиків, інтегрована в мережу пристроїв обробки створює великий потік даних".

Сучасні MicroGrid використовують системи моніторингу здоров'я та використання (Health and usage monitoring systems - HUMS).

Деякими типовими прикладами датчиків, наявних у MicroGrid, є датчики температури, руху, відстані, вологості, звуку, потоку води, газу, дверей. Екологічні датчики використовуються для виявлення взаємодії між користувачем та об'єктами, що допомагає визначити показники щоденної

діяльності людини. До них відносяться датчики, вбудовані в ліжко, стільці, кухонні прилади тощо. Датчики руху виявляють переміщення людини використовуючи оптичні, мікрохвильові, інфрачервоні, акустичні спостереження. Датчики руху зазвичай використовують у складі систем безпеки та освітлення. Також використовується мережа бездротових датчиків руху, де деякі з них об'єднуються з контактними датчиками на дверях.

Інфрачервоні датчики близькості від Motionwireless використовуються системою In-HomeMonitoring (IMS). Ці датчики руху можуть використовуватися в Smart Houses для виявлення безпеки та падіння для людей похилого віку, відстеження користувачів та аналізу поведінкових моделей.

В Університеті охорони здоров’я та науки штату Орегон інфрачервоні датчики руху використовувались для розробки системи реєстрації руху населених пунктів. Ці інфрачервоні датчики розміщені в кожному корпусі мікросітки. Двері також мають магнітні контактні датчики для "відстеження потоку відвідувачів" або для контролю за тим, чи хтось є в будинку.

Однак недоліком є те, що мешканцям потрібно ідентифікувати ідентифікатор радіочастот (RFID), щоб підключитися до одержувача для ідентифікації. Цей недолік також посилювався тим, що в будинку 18 разів було більше однієї людини, кожна людина повинна була бути позначена RFID, щоб можна було дізнатись, де ця людина знаходиться. Кілька об’єктів можна ідентифікувати, одночасно аналізуючи вагу об’єкта.

Ультразвукові датчики також використовуються для виявлення руху. Система Gator MicroGrid використовує ультразвукові датчики для виявлення руху, орієнтації та інформації про місце розташування дорослих людей у MicroGrid. Кілька приймачів дозволяють встановлювати різні відстані для кожного приймача, визначаючи таким чином точне

місцезнаходження. Датчики для визначення камери та руху можуть працювати разом у кількох сценаріях. При реалізації розподіленої мережі інтелектуальних камер, функції якої полягали в локалізації та виявленні падінь користувачів. У системі використовуються малопотужні камери з датчиками класу Mote (вузли сенсорів), створюючи інфраструктуру бездротової мережі.

Крім того, камери використовують децентралізовану процедуру виявлення падінь. Інший приклад використання камер - це ті, які випадково приймають зображення навколишнього середовища.

Зроблені зображення використовуються для виявлення "відповідної інформації": наприклад, якщо людина спить, це відбувається через відсутність руху. Невеликі камери з низькою роздільною здатністю (наприклад, 352-288 пікселів) є кращими в MicroGrid, оскільки вони прості у використанні, вимагають низької роздільної здатності та низької потужності та легко підключаються до джерела живлення для обробки. Крім того, камери використовували децентралізований процес виявлення несправностей.

Датчики для виявлення лихоманки були розроблені для вимірювання температури на основі аналізу термічних зображень. Детектори лихоманки використовують термічну камеру, яку можна розмістити в будь-якому місці MicroGrid (над ліжком, у ванній кімнаті).

Інший проект використовує датчики, поміщені в кімнату для локалізації людини за допомогою Impulse-Radio Ultra-WideBand (IR-UWB).

Базова станція використовувалася для отримання датчиків даних та координат позначки. Використовуючи оцінку, увімкнуту IR-UWB, для оцінки відстані до позначки, використовуючи алгоритми об'ємного часу (RTT).

Бездротові сенсорні вузли (Sensor Nodes - SN) також використовуються в Smart Houses. SN - це невеликі пристрої з

обчислювальними процесорами, блоками бездротового радіо зв'язку та датчиками.

Логічний механізм планування сну на основі кореляції (LCSSM) був впроваджений для зменшення енергоспоживання бездротових SN. В іншому дослідженні був реалізований малопотужний низькочастотний смуговий інтегрований КМОП-фільтр для пасивних інфрачервоних (Passive Infra-red - PIR) датчиків в бездротових SN, що також зменшує споживання електроенергії. PIR також використовувались разом з sensorsflexifore в MicroGrid для виявлення зайнятості людини в об'єктах (диван, туалет, ліжко, стільці тощо). Датчики Flexiforce були реалізовані з динамічним порогом в режимі реального часу, щоб забезпечити більш точне читання вихідного сигналу датчика.

Інші проекти також використовують смартфони, для визначення активності людини за допомогою інтегрованого акселерометра, гіроскопа, GPS та камери.

Акселерометри в наручних годинках використовуються для моніторингу рухової активності. Головна ідея полягає в тому, щоб визначити основні рухи людини (лежати, сидіти, стояти, ходити, бігати, підніматися чи спуститися вниз, працювати на комп'ютері).

Зібраний сигнал надсилається на персональний сервер через радіочастотний зв’язок.

Korel and Koo э розробниками систем про контекстне осмислення з використанням сенсорних мереж тіла (Body Sensor Network - BSN) для постійного моніторингу пацієнтів, щоб виявити аномалії, що загрожують життю. Пацієнт носить або має імплантований пристрій для контролю будьякого фізіологічного параметру (наприклад, артеріального тиску, пульсу). Дослідження зосереджено на контекстно-осмисленому зондування та порівнянні байєсівських мереж, штучних нейронних мереж та прихованих марківських моделей.

Схожий метод контексного осмислення і у цій роботі для виявлення аномальних поведінкових моделей людини у MicroGrid.

1.3. Активність людини у MicroGrid

Люди дотримуються певних моделей у своєму повсякденному житті. У контексті MicroGrid щоденна діяльність користувача створює шаблони, які відіграють важливу роль у прогнозуванні майбутніх подій. Призначення інтелектуального користувальницького інтерфейсу полягає у формуванні MicroGrid як моделі повсякденного життя; Тому мікромережа повинна знаходити повторювані шаблони в діях користувача та передбачати поведінку користувача для подальшої допомоги [8]

.Моніторинг активності користувачів використовується для спостереження та реєстрації дій особи, з метою досягнення цілей комфорту та ефективності, які MicroGrid може запропонувати. Тому необхідно забезпечити здатність до вивчення та застосовування отриманих знань, для адаптації системи керування Microgrid до поведінки користувача. Оскільки користувач створює шаблон, ненормальна поведінка користувачів може бути виявлена шляхом побудови звичайного поведінкового малюнка користувача. Як правило, датчики та фотоапарати в MicroGrid використовуються для відстеження або ідентифікації діяльності користувача та аналізу поведінки людей.

Спостереження за поведінкою користувача використовується для прогнозування його подальшої активності. Дослідження в напрямку виявлення нетипових (аномальних) поведінкових моделей людини є актуальними для систем психофізіологічного моніторингу літніх людей, осіб, що потребують постійного нагляду та піклування, перебувають у постстресових станах, знаходяться під впливом надзвичайних або екстремальних ситуацій.

Таким чином, розробка методів розпізнавання активності людини у MicroGrid грунтується на алгоритмах інтелектуальному керуванні до яких, зокрема відносяться і алгоритми машинного навчання. Алгоритми штучного інтелекту, алгоритми машинного навчання та методи виведення даних використовуються для моделювання та прогнозування поведінки користувача.

З огляду на значні відмінності в поведінковій діяльності між різними людьми та однією і тією ж людиною в різні дні, надзвичайно важливо побудувати схеми тренувань, на основі яких тренується система. Повнота та адекватність навчальних схем є запорукою надійних результатів.

Інформаційна інфраструктура сучасних комплексів MicroGrid оснащена величезною кількістю датчиків, що фіксують різні типи подій і утворюють інформаційну картину функціонування технічних пристроїв та підсистем, а також переміщення та дії людини. Кількість спрацьовувань n двопозиційних датчиків (що фіксують бінарні події типу «вкл/викл») можна розрахувати[11].

Висновки

Наявність людського фактору у MicroGrid призводить до необхідності оперування великими обсягами даних, що змушує дослідників звертатися до спеціалізованих методів роботи з великими даними (Big Data).

Серед сукупності методів машинного навчання найбільш придатним є метод детектування аномалій (Anomaly Detection), який було обрано для визначення нетипових моделей поведінки людини (аномальної поведінки) у MicroGrid.

РОЗДІЛ 2

ОГЛЯД ГІЛКИ ANOMALY DETECTION ТА ІСНУЮЧИХ АЛГОРИТМІВ У MACHINE LEARNING

2.1. Існуючі методи Machine Learning

Інформаційно-керуюча система MicroGrid потребує потужного інструментарію інформаційної підтримки для забезпечення збору обробки та прогнозування великих обсягів різнотипних даних[9].

Важливою частиною таких рішень є модуль аналізу даних, який реалізує алгоритм для виконання процесу аналізу та встановлення зв'язку між подіями. За допомогою цих моделей даних система мікромереж повинна передбачати поведінку користувача на основі історичних даних та розвивати обізнаність про так звані ситуації, тобто розуміти намір користувача у визначений час та змінювати параметри режиму роботи пристроїв та підсистем.

На рис 2.1 зображено класифікаційну схему схема методів Machine Learning.

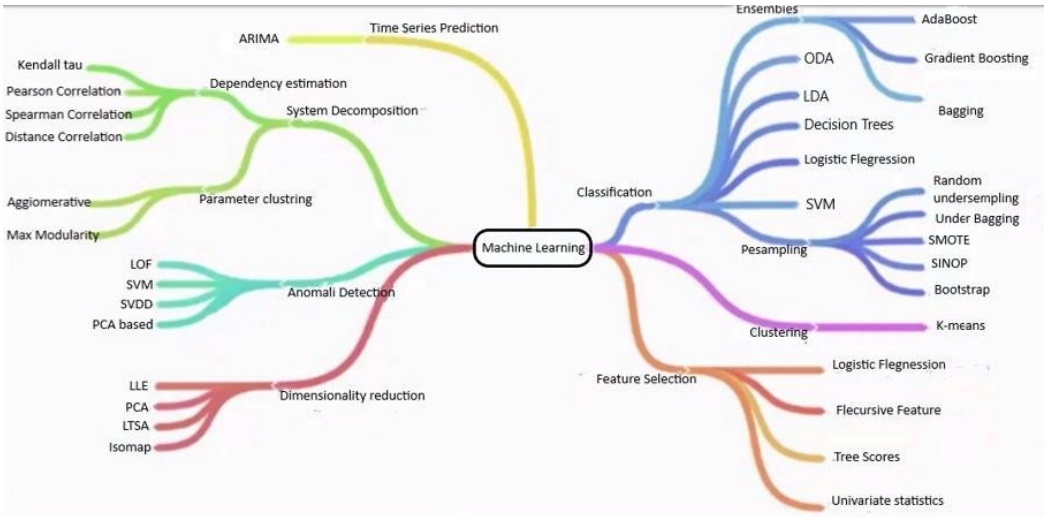

Рис. 2.1 Методи Machine Learning

В даній роботі для формування поставленої мети було обрано гілку Anomaly Detection рис 2.1.

Існуючі методи Anomaly Detection:

1. LOF (Local Outlier Factor);
2. SVM (Support Vector Machine);
3. SVDD (Support Vector Domain Description);
4. PCA (Principal component analysis).

Розглянемо ці методи більш докладно з точки зору їх недоліків та переваг при розв’язанні задачі виявлення аномальної поведінки людини з MicroGrid.

2.2. Огляд існуючих методів Anomaly Detection

LOF (Local Outlier Factor)

Місцевий рівень викидів Маркус М. для пошуку суперечливих точок даних шляхом вимірювання локального відхилення даної точки даних на основі сусідів. Алгоритм виявлення невідповідностей, запропонований у 2000 році Бронігом, Гансом-Пітером Крайгелем, Реймондом Т та Сандером Джорджем

Цей метод заснований на оцінці щільності розміщення об'єктів, що досліджуються на предмет викидів. Предмети в районах з найменшою щільністю трактуються як викиди. Перевага методу LOF перед іншими методами роботи з об'єктами високої щільності полягає в тому, що LOF приймає так звану "локальну щільність". Тому LOF успішно виявляє розбіжності, коли вибірка містить об’єкти різних класів, які не є розбіжностями. На рис 2.2 показаний приклад, коли об'єкти навчальної вибірки належать двом класам C_1 і C_2. Об'єкти в двох класах мають різну щільність. Точки o_1 і o_2, є аномаліями. Завдяки обчисленню локальної щільності класів LOF успішно розпізнає обидві аномалії. Методи,

засновані на обчисленні середньої щільності всіх об'єктів, в більшості випадків виявляють викид o_1 , але пропускають викид o_2.

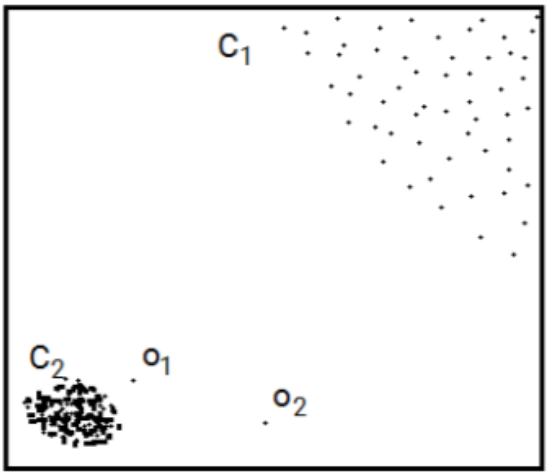

Рис 2.1. Механізм LOF. Випадок областей з різною щільністю

Місцева щільність оцінюється на певній відстані від однієї точки до сусідніх точок. Визначення “доступу”, що використовується в алгоритмі, є додатковим заходом для досягнення більш стабільних результатів у кластерах.

Формальний опис Local Outlier Factor. Нехай $D_k(y)$ є відстанню від точки y до k найближчого сусіда. Досяжністю (reachability distance) точки щодо точки називається величина

$$R_k(x,y) = max(\rho(x,y), D_k(y) \qquad (2.1)$$

Нехай $AR_k(x)$ - середня досяжність точки x щодо k своїх найближчих сусідів, $N_k(x)$ - множина k найближчих сусідів x . Тоді:

$$LOF_k(x) = mean_{y \in N_k(x)} \frac{AR_k(x)}{AR_k(y)} \qquad (2.2)$$

Сенс (2.2) полягає в тому, щоб порівняти середню досяжність точки та її найближчих сусідів. Для “нормальних” даних вірно не тільки, що оцінка (2.1) локальної щільності мала, а й те, що вона незначно відрізняється від такої ж оцінки для найближчих сусідів. приклад роботи алгоритму наведено на рис. 2.2

Переваги Local Outlier Factor. Завдяки розташуванню підходу, алгоритм LOF здатний виявляти викиди в наборі даних, які не можуть випромінюватися в інших областях набору даних. Наприклад, точка на "малій" відстані щільного скупчення є випромінюванням, тоді як точка всередині рідкісного скупчення може мати однакову відстань зі своїми сусідами. Хоча геометричне значення алгоритму можна застосовувати лише до низьковимірних векторних просторів, алгоритм можна застосовувати до будь-якого контексту, де можуть бути визначені відмінності.

Недоліки Local Outlier Factor. Для локальної щільності та коефіцієнтів доступу заборонені встановлені значення, отже, іншими словами, немає чіткого правила про те, що точка буде суперечливою. В одному наборі даних значення 1,1 може означати викиди, в іншому наборі даних і при параметризації (при сильних локальних коливаннях) значення 2 також може знаходитися в області точкової нормальності. Таким чином, дослідження LOF мають високу чутливість для уточнення наборів даних.

SVM (Support Vector Machine)

Метод Support Vector Machines заснований на концепції розподіляючої гіперплощини, які визначають межі рішення. Рівень прийняття рішень - це таке, що відокремлює набір об'єктів, що мають різне членство у класі. Схематичний приклад показаний на рис (2.3). У цьому прикладі об'єкти належать або до класу GREEN або RED. Розмежувальна лінія визначає межу, на правому боці якої всі об'єкти є ЗЕЛЕНИМИ, а ліворуч від яких всі об'єкти КРАСНИМ. Будь-який новий об'єкт (білий круг), що падає праворуч, маркується, тобто класифікується як ЗЕЛЕНИЙ (або класифікується як ЧЕРВОНИЙ, якщо він падає ліворуч від лінії поділу).[16]

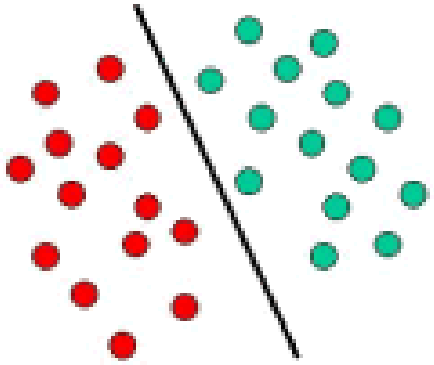

Рис 2.2. SVM - розбиття на класи

Вищезазначене є класичним прикладом лінійного класифікатора, тобто класифікатора, який розбиває набір об'єктів на їх відповідні групи (GREEN і RED в цьому випадку) з лінією. Більшість завдань класифікації, прості, але часто і більш складні структури потрібні для того, щоб зробити оптимальний поділ, тобто правильно класифікувати нові об'єкти (тестові випадки) на основі доступних прикладів . Ця ситуація зображена на рис (2.3). У порівнянні з попередньою схемою, очевидно, що для повного відділення ЗЕЛЕНИХ і ЧЕРВОНИХ об'єктів потрібна крива (яка є більш складною, ніж лінія). Завдання класифікації, засновані на малюванні розділяючих ліній для розрізнення об'єктів різного членства в класі, відомі як класифікатори гіперплощин. Машини підтримки Vector особливо підходять для обробки таких завдань.

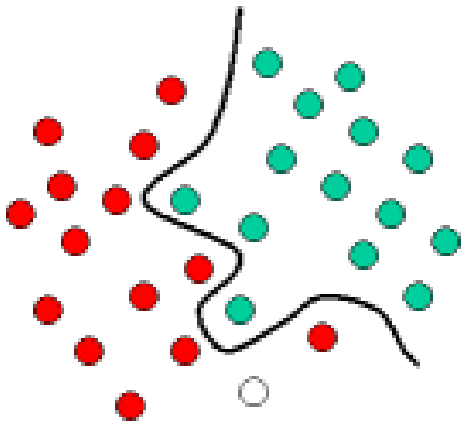

Рис 2.3 Складне розбиття на класи

На рис(2.4) показана основна ідея, що стоїть за SVM. Тут ми бачимо, що оригінальні об'єкти (ліва сторона схеми) відображені, тобто

переставлені, використовуючи набір математичних функцій, відомих як ядра. Процес перегрупування об'єктів відомий як відображення (перетворення). Зауважимо, що в цій новій обстановці відображені об'єкти (права сторона схеми) лінійно відокремлюються і, таким чином, замість побудови комплексної кривої (ліва схема), все, що нам потрібно зробити, це знайти оптимальну лінію, яка може відокремити ЗЕЛЕНІ та КРАСНІ об'єкти.

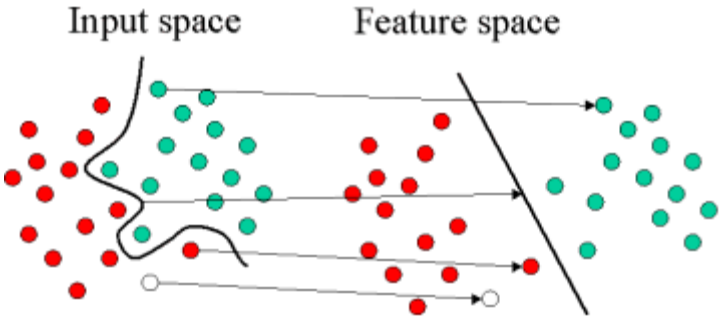

Рис 2.4 Приклад розподілення об’єктів

Support Vector Machine (SVM) - це, перш за все, класичний метод, який виконує завдання класифікації, будуючи гіперплощини в багатовимірному просторі, що розділяє випадки різних міток класу. SVM підтримує завдання регресії та класифікації, а також може обробляти декілька безперервних і категоричних змінних. Для категоричних змінних фіксовану змінну створюють з значеннями випадку як 0 або 1. Таким чином, категоріальна залежна змінна, що складається з трьох рівнів, скажімо (A, B, C), представлена набором з трьох фіктивних змінних:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

Для побудови оптимальної гіперплощини SVM використовує ітеративний алгоритм навчання, який використовується для мінімізації функції помилки. За формою функції помилок SVM-моделі можна розділити на чотири окремі групи:

- Classification SVM Type 1 (також відома як C-SVM classification)
- Classification SVM Type 2 (також відома як nu-SVM classification)

- Regression SVM Type 1 (також відома як epsilon-SVM regression)
- Regression SVM Type 2 (також відома як nu-SVM regression)

Класифікація SVM.

Перший тип класифікації SVM.

Для цього типу SVM навчання передбачає мінімізацію функції помилки:

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

з урахуванням обмежень:

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots, N$$

де C - константа ємності, w - вектор коефіцієнтів, b - постійна, і представляє параметри для обробки нероздільних даних (входів). Індекс і позначає N випадків навчання. Зауважимо, що представляє мітки класів і хі представляє незалежні змінні. Ядро використовується для перетворення даних з вхідного (незалежного) до простору ознак. Слід зазначити, що чим більше C, тим більше помилка покарається. Таким чином, C слід вибирати обережно, щоб уникнути надмірного прилягання.

Другий тип класифікації SVM.

На відміну від класифікації SVM Type 1, модель класифікації SVM типу 2 мінімізує функцію помилок:

$$\frac{1}{2} w^T w - \nu\rho + \frac{1}{N} \sum_{i=1}^{N} \xi_i$$

з урахуванням обмежень:

$$y_i \left(w^T \phi(x_i) + b \right) \geq \rho - \xi_i, \xi_i \geq 0, i = 1, \ldots, N \text{ and } \rho \geq 0$$

У регресійному SVM необхідно оцінити функціональну залежність залежної змінної y від набору незалежних змінних x. Вона передбачає, як і інші проблеми регресії, що взаємозв'язок між незалежними і залежними

змінними задається детермінованою функцією f плюс додавання деяких адитивних шумів:

Регресія SVM.

$$y = f(x) + \text{шум}$$

Задача полягає в тому, щоб знайти функціональну форму для f, яка може правильно передбачити нові випадки, які SVM не були представлені раніше. Це може бути досягнуто шляхом навчання моделі SVM на наборі зразків, тобто навчального набору, процесу, який включає, наприклад, класифікацію (див. Вище) послідовну оптимізацію функції помилки. Залежно від визначення цієї функції помилки можна розпізнати два типи моделей SVM:

Перший тип регресійного SVM.

Для цього типу SVM функція помилки:

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i + C\sum_{i=1}^{N} \xi^*_i$$

ми мінімізуємо:

$$w^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i^*$$
$$y_i - w^T \phi(x_i) - b_i \le \varepsilon + \xi_i$$
$$\xi_i, \xi_i^* \ge 0, i = 1, \ldots, N$$

Другий тип регресійного SVM.

Для цієї моделі SVM функція помилки задається:

$$\frac{1}{2} w^T w - C\left(v\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \left(\xi_i + \xi_i^*\right)\right)$$

ми мінімізуємо:

$$\left(w^T \phi(x_i) + b\right) - y_i \le \varepsilon + \xi_i$$
$$y_i - \left(w^T \phi(x_i) + b_i\right) \le \varepsilon + \xi^*_i$$
$$\xi_i, \xi^*_i \ge 0, i = 1, \ldots, N, \varepsilon \ge 0$$

Існує декілька ядер, які можна використовувати в моделях підтримки векторних машин. До них відносяться лінійна, поліноміальна, радіальна базисна функція (RBF) і сигмоподібна:

Функції ядра.

$$K(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} \mathbf{X}_i \cdot \mathbf{X}_j & \text{Linear} \\ (\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C)^d & \text{Polynomial} \\ \exp(-\gamma |\mathbf{X}_i - \mathbf{X}_j|^2) & \text{RBF} \\ \tanh(\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C) & \text{Sigmoid} \end{cases}$$

Де $K(\mathbf{X}_i, \mathbf{X}_j) = \phi(\mathbf{X}_i) \cdot \phi(\mathbf{X}_j)$ тобто функція ядра, представляє точковий продукт вхідних даних, перетворених у просторі характеристик вищого розміру.

Ширина розділяючої смуги. Щоб розділяюча гіперплощина якнайдалі відстояла від точок вибірки, ширина смуги повинна бути максимальною. Нехай x_- і x_+ - дві довільні точки класів -1 і +1 відповідно, що лежать на межі смуги. Тоді ширина смуги є:

$$\left((x_+ - x_-)\frac{w}{||w||}\right) = \frac{(w,x_+)-(w,x_-)}{||w||} = \frac{(w_0+1)-(w_0-1)}{||w||} = \frac{2}{||w||}$$

Ширина смуги максимальна, коли норма вектора w мінімальна. Отже, в разі, коли вибірка лінійно роздільна, досить прості геометричні міркування приводять до наступної задачі: потрібно знайти такі значення параметрів w і w_0 , при яких норма вектора w мінімальна. Це задача квадратичного програмування. Вона буде детально розглянута в наступному розділі. Потім буде зроблено узагальнення на той випадок, коли лінійної роздільності немає.

SVDD (Support Vector Domain Description)

З набору даних, що містить N об'єктів даних $\{x_i, i = 1, \ldots, N\}$, потрібний опис. Ми Виводимо сферу з мінімальним обсягом, що містить всі (або більшість) об'єктів даних. Це дуже чутливо до найбільш

віддаленого об'єкта в цільовому наборі даних. Коли один або кілька дуже віддалених об'єктів знаходяться отримано дуже велику сферу які не будуть представляти дані дуже добре.

Таким чином, ми допускаємо деякі дані за межами сфери і вводять слабкі змінні. З сфери, описуваної центром a і радіусом R, ми мінімізуємо радіус:

$$F(R, a, \xi_i)=R^2 + C \sum_i \xi_i, \quad (2.1)$$

де змінна C дає середнє значення між простою (або обсягом сфери) і кількість помилок (кількість відхилених цільових об’єктів).

Це повинно бути мінімізовано під обмеження

$$(x_i - a)^T (x_i - a) \leq R^2 + \xi_i \quad \forall_i, \xi_i \geq 0. \quad (2.2)$$

Включаючи ці обмеження в (1), будуємо рівняння за методом Лагранжа:

$$L(R, a, a_i, \xi_i = R^2 + C \sum_i \xi_i - \sum_i a_i\{R^2 + \xi_i - (x_i^2 - 2ax_i + a^2)\} - \sum_i \gamma_i \xi_i, \quad (2.3)$$

з множниками Лагранжа $a_i \geq 0$ і $\gamma_i \geq 0$. Встановивши часткові похідні на 0, отримаємо нові обмеження:

$$\sum_i a_i = 1, a = \frac{\sum_i a_i x_i}{\sum_i a_i} = \sum_i a_i \, x_i,$$

$$C - a_i - \gamma_i = 0 \ \forall_i \quad (2.4)$$

Оскільки $a_i \geq 0$ та $\gamma_i \geq 0$, то можна вилучити змінні γ_i з третього рівняння в (4) і використовувати обмеження $0 \leq a_i \leq C. \quad \forall_i$.

Переписування рівняння. (2.3) і повторно встановлюють рівняння. (2.4)

дати максимізацію по відношенню до a_i:

$$L = \sum_i a_i \, (x_i \cdot x_i) - \sum_{ij} a_i \, a_j (x_i \cdot x_j) \quad (2.5)$$

з обмеженнями $0 \leq a_i \leq C, \sum_i a_i = 1$.

Друге рівняння в (2.4) говорить про те, що центр сфери є лінійною комбінацією об'єктів даних, з ваговими коефіцієнтами a_i, отриманими шляхом оптимізація рівняння. (2.5). Тільки для невеликого набору об'єктів

рівність у формулі. (2.2) задовольняється: це об'єкти, що знаходяться на межі сфери себе.

Для цих об'єктів співвідношення аі будуть ненульові і називатимуться об'єктами підтримки. Тільки ці об'єкти необхідні в описі сфери. Можна отримати радіус R сфери шляхом обчислення відстані від центру сфери до вектора підтримки з масою менше ніж C. Об'єкти, для яких аі В Є С потрапили верхня межа в (2.4) і знаходяться за межами сфери. Ці вектори підтримки вважаються викидами.

Щоб визначити, чи знаходиться контрольна точка z в межах сфера, відстань до центру сфери для розрахунку. Тестовий об'єкт z приймається, коли ця відстань менше, ніж радіус, тобто коли $(z-a)^T(z-a) \leq R^2$.. Висловлюючи центр сферу в термінах допоміжних векторів ми приймаємо об'єкти, коли

$$(z \cdot z) - 2\sum_i a_i\,(z \cdot x_i) + \sum_{ij} a_i\,a_j(x_i \cdot x_j) \leq R^2. \quad (2.6)$$

Узагальнення до інших ядер. Простий метод обчислює лише область навколо даних у вхідному просторі. Як правило, дані не поширюються сферично, якщо віддалені об'єкти ігноруються. Отже, загалом, ми не можемо довіряти отримання дуже суворого опису. Оскільки проблема повністю описана з точки зору внутрішніх добутків між векторами (2.5) і (2.6), метод може бути додатково розроблений. [15]

Добутки $(x_i \cdot x_j)$ можуть бути замінені на функцію ядра $K(x_i \cdot x_j)$, коли це ядро $K(x_i \cdot x_j)$ задовольняється Теорема Мерсера[17]. Це неявно відображає об'єкти x_i в деякому просторі функцій і коли підходить вибирається простір ознак, кращий, більш жорсткий опис можна отримати. Немає явного відображення.

Потрібна проблема повністю виражена з точки зору $K(x_i \cdot x_j)$.Тому замінюємо всі внутрішні продукти $(x_i \cdot x_j)$ належним $K(x_i \cdot x_j)$. Дані опису доменів тепер задаються рівнянням (див. (2.5))

$$L = \sum_i a_i K(x_i \cdot x_j) - \sum_i a_i a_j K(x_i \cdot x_j) \quad (2.7)$$

з обмеженнями $0 \leq a_i \leq C$, $\sum_i a_i = 1$. Тестовий об'єкт z приймається, коли (див. (2.6):

$$K(z,z) - 2\sum_i a_i (z \cdot x_i) + \sum_{ij} a_i a_j (x_i \cdot x_j) \quad (2.8)$$

Різні функції ядра K призводять до розрізнення меж опису у вихідному просторі введення. Проблема полягає в тому, щоб підібрати відповідну функцію ядра $K(x_i \cdot x_j)$. Обговорюються два варіанти: поліноміальне ядро і ядро Гауса.

Найпершим вибором для ядра $K(x_i \cdot x_j)$- є розширений добуток: $K(x_i \cdot x_j)$=$(x_i \cdot x_j + 1)^d$, де вільний параметр d – ступінь поліноміального ядра.

Це ядро відображає об'єкти у просторі просторових об'єктів, додаючи продукти оригінальні риси, до ступеня d. (Наприклад, а 2D вектор $(x_{1,}x_2)$відображається в $(x_{1,}x_2x_1x_2, x_1^2x_2^2)$ коли використовується поліноміальне ядро з $d = 2$. Це ядро взагалі не призводить до хорошого жорсткі описи. Для більш високих ступенів d, значення об'єктів, найбільш віддалених від походження система координат збільшується і переповнює все інші внутрішні продукти. Цей ефект показаний на рис з двовимірним набором даних, що містить 10 об'єктів. Для різних значень ступеня ($d = 1$,$d = 5$,$d = 25$) обчислюється опис сфери.

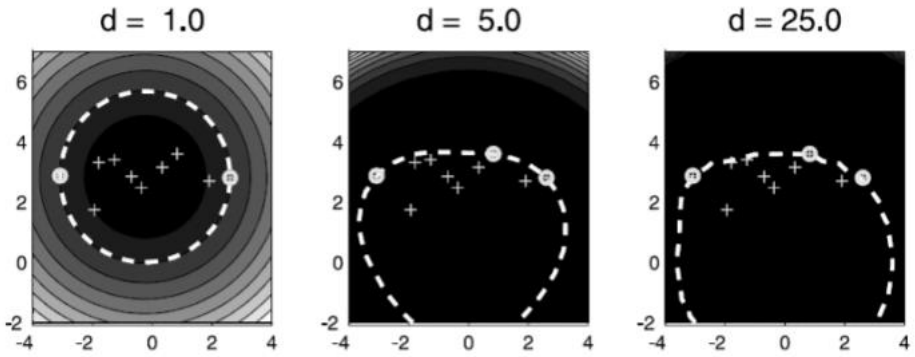

Рис 2.5. Сфери при різних значеннях ступеня

Наведено відстань до центру сфери у вихідному просторі введення. Пунктирна біла лінія перетинання опорних векторів (позначено Малі кола) є кордоном опису.

Найдальші об'єкти знаходяться у верхній частині оригіналу, і хоча ці об'єкти не є найвіддаленішими об'єктами у даних у двох вимірах, вони стають опорними векторами, коли використовується вищий ступінь. Зверніть увагу, що приймаються більшість місць в'їзду. Це опис результатів дуже великої та розрідженої площі вихідної двовимірної точки входу.

Для зменшення зростаючих відстаней для більших просторових ознак, гауссовське ядро $K_G(x_i, x_j) = exp(-(x_i - x_j)^2/s^2$ є більш доцільним. Рівняння (2.7) набуває вигляду:

$$L = 1 - \sum_i a_i - \sum_{i \neq j} a_i \, a_j K_G(x_i, x_j) \quad (2.9)$$

а рівняння (2.8) -

$$-2 \sum_i a_i K_G(z, x_i) \leq R^2 - C_x - 1 \quad (2.10)$$

де C_x залежить тільки від опорних векторів та a_i, а не від тестового об'єкта z.

На рис. 2, знову ж таки, показано 2D набору даних, що містить 10 об'єктів. Тепер вектор підтримки опису домену з гауссовим ядром для використовуються різні значення S. Параметр ширини s коливається від дуже малого (S = 1.0 у лівій частині))Gure) до великих (S = 1.0 у правих краях). Зауважимо, що кількість підтримуючих векторів зменшується і що опис стає більш подібним до сфери.

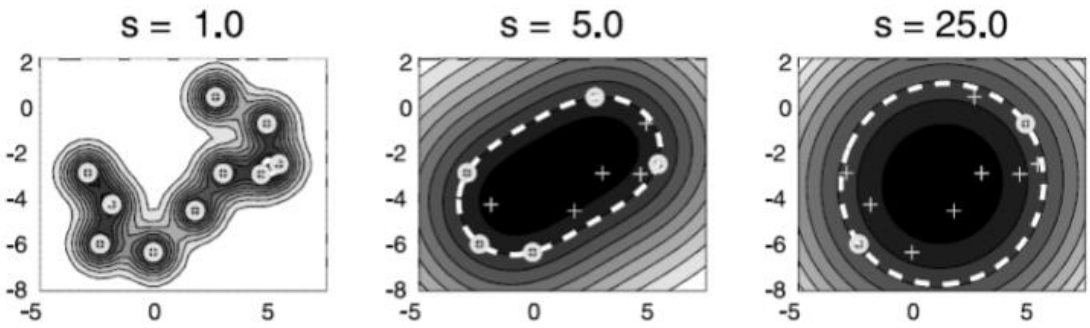

Рис 2.6. Сфери при різних значеннях ширини

Можна вивести явні рішення для рівняння. (7) для дві різні екстремальні ситуації, одна для дуже малі значення і одне для дуже великих значень s. Для дуже малих , $K_G(x_i, x_j) \simeq 0$, $i \neq j$ та $L = 1 - \sum_i a_i^2$.. Це максимально, коли $a_i = 1/N$ і L становить $1 - 1/N$. Це схоже на Парзен оцінка щільності, де кожен об'єкт підтримує а ядро (див. (10)). Всі відстані до центру сфера стає $1 - 1/N$.

Для дуже великих $S, K_G(x_i, x_j) = 1$ і $L = 1 - \sum_i a_i^2 - \sum_{i \neq j} a_i\, a_j$. Це максимально, коли всі $a_i = 0$, за винятком одного $a_i = 1$, і всіх відстаней до центром сфери стають 0. Це ввже досить велика сфера не буде отримано на практиці і не буде достатньо велика, щоб дати рівний $K_G(x_i, x_j)$ для всіх пар i, j.

У самому правому підводі на рис. 2 реалістично Наведена гранична ситуація. Опис даних знову найменша сфера, яка охоплює повний набір даних, без викидів. Розширення Тейлора рівняння. (9) показує, що коли ігноруються більш високі порядки, (34) (5) виходить (до масштабування і коефіцієнт вибору).

У випадку середніх значень S (середній підграфік на рис. 2.8) лише частина об'єктів стає опорними об'єктами. Вираз (2.24) показує, що в цьому випадку отримана відредагована та зважена оцінка щільності Парцена. Параметр C дає верхню межу для параметрів a_i, таким чином, обмежує опорні вектори (2.6). Коли об'єкт x_i отримує $a_i = C$, опис більше не буде

адаптовано до цього об'єкта, і він залишиться поза сферою. Через обмеження $\sum_i a_i = 1$ і $a_i \geq 0$ - лише ті варіанти, для яких C може мати будь-яке входження до рішення рівняння. (2.7), коли $1/N \leq C \leq 1$. Для $C \leq 1/N$ неможливо знайти рішення, тому тоді обмеження $\sum_i a_i = 1$ ніколи не може бути виконано, тоді як для $C > 1$ завжди можна знайти вирішення (a_i завжди менше або дорівнює 1). Коли C обмежується малими значеннями, перевага виходу за межі сфери не є дуже великою, а більша частина об'єктів може бути поза сферою. На практиці перевага C не дуже критична.

Узагальнення. Щоб отримати вказівку узагальнення або перевищуючи характеристики СВДД, ми повинні отримати вказівку (2.1) кількості цільових шаблонів, які будуть відхилені (помилки першого роду) цим описом і (2.2) кількістю віддалених шаблонів, які будуть прийняті (помилки другого роду).

Ми можемо оцінити похибку застосування методу першого типу відпустки для вивчення набору, що містить цільовий клас. Коли ми знімаємо опору з навчального об’єкта, буде знайдено оригінальне рішення, і всі об’єкти навчання будуть знайдені. Коли елемент підтримки виключається, опис оптимального обсягу можна скоротити, оскільки цей елемент підтримки знаходиться в межах діапазону. Цей занедбаний предмет буде відхилено, а інші навчальні предмети все одно прийматимуться (оскільки метод вивчає дані з них). Тому про це можна зробити висновок помилково.

$$E[P(error)] = \frac{\#SV}{N} \qquad (2.11)$$

де $\#SV$- це кількість опорних векторів.

Коли ми використовуємо ядро Гауса, ми можемо регулювати кількість векторів підтримки шляхом зміни параметр ширини s. Тому ми можемо також встановити помилку першого роду. Коли кількість векторів

підтримки занадто великі, нам потрібно збільшити s, в той час, коли номер занадто низький, ми повинні зменшення s. Щоб перевірити, наскільки добре оцінюється рівняння. (2.11), ми нанесли на рис. 3 оцінку помилки першого роду як функції від параметр ширини s. Метод був застосований до а двовимірний набір даних, що містить 10 об'єктів. Також похибка, оцінена на незалежному тесті наведено 100 об'єктів. Можна зробити висновок ця оцінка добре працює.

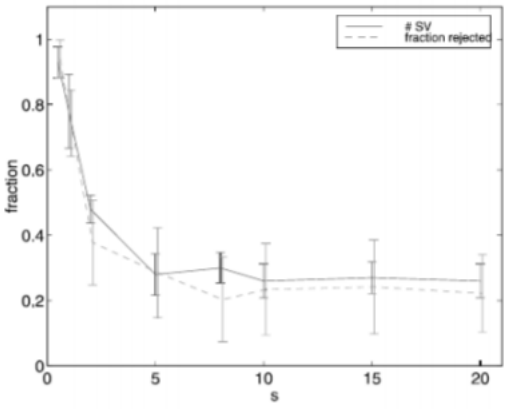

Рис. 2.7. Порівняння частки об'єктів та оцінка помилок

Отже, коли потрібний опис набору даних, ми можемо заздалегідь встановити обмеження на очікуване коефіцієнт відхилення цільових даних. Лагранжиан від рівняння. (2.9) вирішується і очікувана помилка для цей розчин отримують за допомогою рівняння. (2.11). При цьому помилка занадто велика, параметр ширини s збільшується, або коли ця помилка може збільшитися, Ширина параметра s зменшується. Це гарантує параметр ширини в SVDD адаптований для проблеми під рукою, враховуючи помилку.

Можливість прийому об'єктів, далеких від опису місцевості, другого типу помилок, не може бути оцінена за цим показником. Як правило, лише хороший опис мети заняття доступний як навчальний набір. Усі інші моделі вважаються викидами. Щоб отримати оцінку другого типу помилок, дані повинні бути створені та перевірені навколо вихідного набору даних. Цей метод вимагає методу отримання або створення даних навколо

навчальних даних, але не в навчальному наборі. Крім того, кількість тестів моделей має бути достатньою для правильної оцінки, що може бути проблемою у просторових просторових об'єктах.

Ми вирішуємо цю проблему, використовуючи класифікаційні задачі тестового методу. З класифікаційної функції ми беремо один клас як клас переміщення, а всі інші класи використовуватимуться як цільовий клас. Таким чином можна створити художні викиди. Це означає, що партійне впровадження розпочато. Ці проблеми класифікації часто мають класи, що перекриваються, і при використанні проблем класифікації продуктивність вихідних методів буде, таким чином, меншою, ніж звичайні методи класифікації для класифікаційної роботи. Однак це дає вказівку на ефективність при порівнянні різних методів викидів.

Метод головних компонент PCA

Метод головних компонент (Principal component analysis) використовується для перетворення даних в стек з вхідного багатовимірного атрибутивного простору в новий багатовимірний атрибутивний простір, осі якого повернуті по відношенню до осей вихідного простору. Осі (атрибути) в новому просторі некорельовані.

Основною метою перетворення даних при аналізі методу основних компонентів є стиснення даних шляхом усунення існуючої надмірності, пошук взаємопов'язаних компонентів, вибір основних компонентів та оцінка відмінностей між кожним з них.

Перший основний компонент буде характеризуватися найбільшими варіаціями, другий варіант буде відповідати другому за величиною значенням. У більшості випадків перші три чи чотири компоненти аналізу основних компонентів описують різницю понад 95 %%. Решту можна викинути. Оскільки новий набір містить менше компонентів, а мінливість

вихідного набору на 95% залишається незмінною, обчислення виконуватимуться швидше, зберігаючи їх точність.

Для інструменту PCA потрібно, щоб були визначені вхідні канали, число головних компонент, в які будуть трансформуватися дані, ім'я вихідного файлу статистики і ім'я вихідного набору.

Основні принципи аналізу за методом головних компонент. При використанні двоканального набору переміщення і обертання осей і трансформація даних здійснюється наступним чином:

• дані наводяться на діаграмі розсіювання.

• для зв'язку точок на діаграмі розсіювання обчислюється еліпс (рис. 2.8).

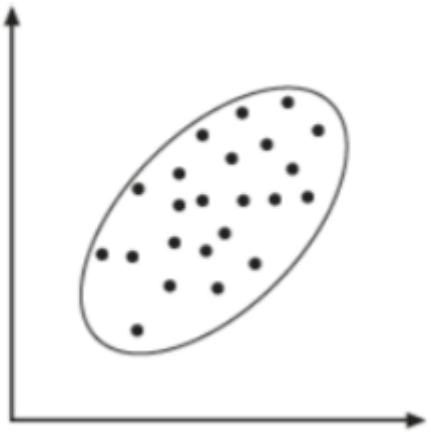

Рис 2.8. Побудована межа еліпса

Визначається головна вісь еліпса (рис. 2.9). Головна вісь стане новою віссю x - першою головнною компонентою (PC1). PC1 має найбільшу дисперсію, тому що це найбільший розріз, який можна зробити через еліпс. Напрямком PC1 є власний вектор, а його величиною - власне значення. Кут осі x до PC1 - це кут повороту, який використовується в трансформації.

Далі обчислюється перпендикуляр до PC1. Ця лінія є другою головною компонентою (PC2) і новою віссю для вихідної осі y (рис. 2.10).

Рис 2.9. Перша головна компонента

Нова вісь відображає найбільшу дисперсію після PC1.

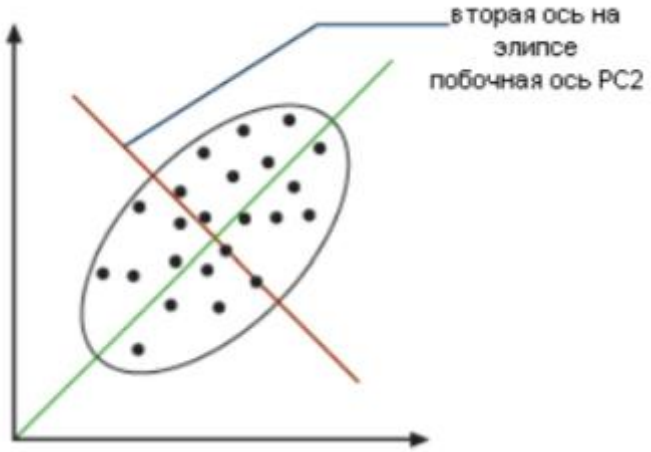

Рис. 2.10 Друга головна компонента

При використанні власних векторів, власних значень і обчисленої коваріаційної матриці вхідних даних багатоканального набору створюється лінійна формула, яка визначає зрушення і поворот. Ця формула застосовується для трансформації кожного значення відносно нової осі.

Завдання аналізу головних компонент має, як мінімум, чотири базових версії:

• апроксимувати дані лінійними залежностями меншої розмірності;

• знайти підпростір меншої розмірності, в ортогональній проекції на який розкид даних (тобто середньоквадратичне відхилення від середнього значення) максимальний;

• знайти підпростір меншої розмірності, в ортогональноій проекції на який середньоквадратична відстань між точками максимальна;

• для досліджуваної багатовимірної випадкової величини побудувати таке ортогональное перетворення координат, в результаті якого кореляції між окремими координатами будуть дорівнювати нулю. Перші три версії оперують кінцевими множинами даних. Вони еквівалентні і не використовують жодної гіпотези про статистичний породженні даних. Четверта версія оперує випадковими величинами. Кінцеві множини розглядаються як вибірки з даного розподілу, а рішення трьох перших завдань - як наближення до розкладання по теоремі Кархунена - Лоєва («істинного перетворення Кархунена - Лоєва»).

2.3 Порівняння існуючих алгоритмів обробки даних

Існує багато методів детектування аномалій, їх поділяють на методи навчання з учителем та без нього , так як в нашій роботі описано метод навчання з учителем важливим при доведенні методу буде порівняння його з іншими методами навчання з учителем а також , методами навчання без нагляду.

Метод навчання з учителем. Методи навчання з учителем вимагали маркованого навчального набору, що містить як нормальні так і аномальні зразки даних, для побудови прогностичної моделі. Теоретично методи навчання з учителем забезпечують більш високий рівень виявлення, ніж методи без нагляду, оскільки вони мають доступ до додаткової інформації. Однак існують деякі технічні питання, які роблять ці методи неточними, як вони повинні бути. Перша проблема полягає у нестачі набору навчальних даних, що охоплює всі області. Більш того, отримання точних міток є

проблемою, і навчальні комплекти зазвичай містять деякі шуми, які призводять до більш високих частот помилкових тривог. Найбільш поширені алгоритми навчання з учителем, Supervised Neural Networks, Support Vector Machines(SVM), k-Nearest Neighbors, Bayesian Networks and Decision Tree.

Метод навчання без нагляду. Вони ґрунтуються на двох основних припущеннях. По-перше, вони припускають, що більшість даних є звичайним потоком і лише дуже малий відсоток є аномальним. По-друге, вони передбачають, що аномальні дані статистично відрізняються від звичайних даних. Згідно з цими двома припущеннями, групи даних, які часто з'являються, вважаються нормальними, тоді як випадки, які значно відрізняються від більшості випадків, вважаються аномальними. Найбільш поширені методи без нагляду, K-Means, Self-organizing maps (SOM), C-means, Expectation-Maximization Meta algorithm (EM), Adaptive resonance theory (ART), Unsupervised Niche Clustering (UNC) and One-Class Support Vector Machine.

На основі попередніх описів різних методів виявлення розбіжностей у таблиці (2.1) порівнюються найпоширеніші методи. Порівняння узагальнює плюси і мінуси кожного. Слабкі сторони методу ідентифікації бази знань. Методи та методи навчання без нагляду входять до складу вчителя для виявлення розбіжностей. Порівняння показало, що метод викладання з викладачем є важливішим, ніж метод викладання без нагляду, якщо у даних тестів немає невідомих розбіжностей. Між методами викладання вчителів відмінні показники досягаються за допомогою нелінійних методів, таких як SVM.

Таблиця 2.1.

Переваги та недоліки алгоритмів обробки даних

Метод	Переваги	Недоліки
K -Nearest Neighbor	• Дуже легко зрозуміти, коли є декілька перемінних. • Корисно для побудови моделей, які включають нестандартні типи даних, наприклад, текст.	• Чутливий до вибору функції подібності, яка використовується для порівняння даних. • Відсутність принципового способу вибору k, за винятком крос-перевірки або подібних. • Дорога обчислювальна техніка техніка.
Neural Network	• Нейронна мережа може виконувати завдання, які неможливо виконати лінійною програмою. • Коли елемент нейронної мережі виходить з ладу, він може продовжуватися без будь-яких проблем з їх паралельною природою. • Нейронна мережа навчається і не потребує перепрограмування. • Вона може бути реалізована в будь-якій програмі.	• Для роботи нейронної мережі потрібна підготовка тестових даних. • Архітектура нейронної мережі відрізняється від архітектури мікропроцесорів, тому потребує емуляції. • Необхідний високий час обробки для великих нейронних мереж.
Decision Tree	• Простий для розуміння та інтерпретації. • Необхідна невелика підготовка даних • Здатний обробляти як числові, так і категоріальні дані. • Використовує модель white box. • Можлива перевірка моделі з використанням статистичних тестів. • Надійна. • Виконайте роботу з великими даними за короткий час.	• Проблема навчання оптимального дерева рішень, як відомо, є NP-повною за декількома аспектами оптимальності і навіть для простих понять. • Учні, які приймають рішення, створюють надскладні дерева, які не зовсім добре узагальнюють дані. • Є концепції, які важко вивчити.
Support Vector Machine	• Знаходить оптимальне розділення гіперполя. • Може працювати з великими даними (BigData). • Деякі ядра мають нескінченний вимір Вапник-Червоненкіс, що означає, що вони можуть вивчати дуже складні концепції.	• Потрібні як позитивні, так і негативні тестові дані. • Потрібно вибрати хорошу функцію ядра. • Потрібно багато пам'яті та процесорного часу.

Self-organizing map	• Простий і легкий для розуміння алгоритм, який працює. • Алгоритм, який працює з нелінійним набором даних. • Відмінна можливість візуалізувати висовикомірні дані на 1 або 2-мірному просторі робить її унікальною, особливо для зменшення розмірності.	• Алгоритм, що займає багато часу
K-means	• Низька складність.	• Необхідність визначення параметру k. • Чутливі до точок даних шуму та викидів. • Кластери чутливі до початкового значення центроїдів.
Fuzzy C-means	• Дозволяє точці даних бути в декількох кластерах. • Більш точне уявлення проприроду даних.	• Потрібно визначити c, число кластерів. • Необхідно визначити значення відсічення даних. • Кластери чутливі до початкового призначення центроїдів.
Expectation-Maximization Meta	• Можна легко змінити модель для адаптації до різного розподілу наборів даних. • Число параметрів не збільшується з підвищенням навчальних даних.	• Повільна конвергенція в деяких випадках.

Висновки

Огляд чотирьох методів машинного навчання показав, найбільши доцільним для використання є метод Support Vector Machine (SVM). В методі SVM наявні тренувальні дані (класу) для порівняння з тестовими даними, що надходять у систему. На відміну від методу Local Outlier Factor, де використовується поняття щільності, метод SVM порівняння та розподілення даних використовує гіперплощини, що є більш наближеним до задачі детектування аномальної поведінки. Метод PCA хоч і дозволяє для пошуку аномалій зменшити дані, але може призводити до значних похибок. На відміну від PCA, метод SVM дозволяє чітко розділити гіперплощини та знайти аномальні данні в виборці. Метод SVDD не є

доцільним для використання він є більш складною похідною від SVM і використовує чотиривимірну модель, що утворює додаткові ускладнення, які не є доцільними в рамках розв’язання задачі дипломної роботи, тому для подальших досліджень було обрано метод SVM. Серед існуючих алгоритмів машинного навчання проведено порівняльний аналіз та виявлено що для можливості використання в системах MicroGrid також найкраще підходить метод SVM. Так як він найкраще підходить для обробки великих даних які присутні в системі MicroGrid, також він найкраще обробляє однотипні данні, наприклад ВКЛ/ВИКЛ датчиків які обробляють системні процеси.

РОЗДІЛ 3

Прогнозуання навантаження на основі машинного навчання

3.1 Рівень попиту на електроенергію в системах MicroGrid

Децентралізовані оператори енергетичних систем, агрегатори, постачальники, менеджери та інші зацікавлені сторони зазнають труднощів через кілька конфліктів, що варіюються від недостатнього постачання електроенергії до зростаючого споживання.

Криві попиту на електроенергію в децентралізованих енергетичних системах (таких як будівлі, енергетичні спільноти, мікромережі, віртуальні електростанції, місцеві енергетичні мережі тощо) відрізняються від типових кривих попиту на електроенергію, які представляють споживання електроенергії в масштабах всієї країни чи регіону.

Це робить традиційні методи (розроблені для національних або регіональних прогнозів попиту на електроенергію) непридатними для їх прямого застосування в децентралізованих енергетичних системах із двох чітких причин. У децентралізованих енергетичних системах не тільки загальний рівень попиту на електроенергію в рази менше, ніж регіональний або національний рівень попиту, але й профіль попиту на електроенергію виявляє більші коливання і, як правило, не відповідає одному і тому ж профілю.

Тому нещодавнє впровадження децентралізованих енергетичних систем вимагає відповідних та застосовних інструментів вибору функцій (FS) та моделей прогнозування для економічного та ефективного моделювання споживання.

Вибір ознак - це процедура вибору підмножини найважливіших ознак (атрибутів, змінних чи предикторів) для використання у розробці прогнозної моделі.

Базові знання про статистичні моделі можуть допомогти зрозуміти ефективний метод вибору ознак, важливий для остаточного результату прогнозування моделей прогнозування. Однак ефективний підхід до вибору ознак повинен бути належним чином розроблений, впроваджений та протестований для конкретного програмного забезпечення, про яке йдеться. У сучасну епоху "великих даних" набори даних наповнені величезними даними, зібраними з апаратів і датчиків. Це робить дані великими розмірами, і стало дуже поширеним спостереження за наборами даних із сотнями (навіть тисячами) змінних.

Те саме стосується енергетичного сектору, оскільки концепція розумних мереж була розроблена та впроваджується на основі ідеї IoT та складної взаємодії даних між різними зацікавленими сторонами.

Коли дані представлені з дуже великою розмірністю, моделі прогнозування, як правило, задихаються, оскільки:

1. Час навчання, перевірки та тестування збільшується в геометричній прогресії із збільшенням кількості змінних.
2. Моделі прогнозування матимуть зростаючий ризик переобладнання із збільшенням кількості предикторів.
3. Точність прогнозування зіткнеться із сукупною загрозою зменшення із збільшенням кількості функцій.

Тому вибір функцій є дуже важливим елементом у процесі роботи вчених-вчених. Різні дослідницькі групи застосовували різні методи вибору функцій для різних застосувань та сценаріїв. Однак дуже мало з них поєднали та дослідили засоби вибору особливостей та моделі прогнозування. Більше того, наразі не існує стандартного та загально узгодженого методу вибору ознак.

Дослідження та розробки з метою пошуку найбільш ефективних інструментів вибору функцій досі тривають різними незалежними дослідницькими групами та установами. Ефективні та адаптивні методи

вибору ознак повинні розроблятися, впроваджуватися та тестуватися для бажаних додатків та сценаріїв постійно, оскільки постійно з’являються нові джерела даних, політики та алгоритми. Це основні причини, які спонукали розробку та впровадження методу.

Стратегії вибору функцій важливі для таких основних переваг:

1. Скоротіть час обчислення
2. Зменшити вимоги до зберігання даних
3. Спростити моделі, зробивши їх зручними для користувача
4. Покращити зрозумілість даних

Основним аргументом при застосуванні методу є те, що вихідний набір даних містить деякі змінні, які або дублюються, або є аномаліями в системі, і тому їх можна усунути, не спричиняючи великої шкоди інформації. Кілька досліджень довели, що надлишкові та нерелевантні особливості знижують точність та можливості узагальнення прогнозних моделей.

Ось чому в наш час дослідження великих данних стали дуже популярними в галузі штучного інтелекту, машинного навчання (ML), глибокого навчання (DL) та статистики.

Різні відповідні дослідження демонструють, що загальні витрати енергії (виробництво енергії, експлуатація та закупівлі) можуть бути значно зменшені, застосовуючи концепцію реагування на попит в системах MicroGrid.

Однак методи прогнозування попиту на електроенергію в цих системах не досліджені належним чином. Найбільш застосовні методи обмежені для прогнозування величини попиту на електроенергію у великих масштабах (загальнонаціональних або регіональних) та ігнорують конкретні потреби в електроенергії для менших підприємств, таких як будівлі, енергетичні спільноти, віртуальні електростанції, локальні енергетичні мережі, тощо,

які мають рівнозначне значення для оптимізації енергосистеми, стійкості та ефективності.

Таким чином, метою даної роботи є запропонувати та впровадити підхід до вибору особливостей для моделювання та прогнозування коливального попиту на електроенергію в системах MicroGrid загалом та будинках зокрема. Результати допоможуть зацікавленим сторонам розподіленої енергетичної системи ефективно використовувати обмежені енергетичні ресурси та регулювати виробництво та гнучкий рівень попиту.

Точність прогнозування в основному була неодмінною метою досліджень прогнозування. Точність моделей прогнозування покладається не тільки на конфігурації моделей та пов'язані з ними методи навчання, а й на область предикторів, яка встановлюється за допомогою початкового простору аномальних данних в системі.

В основному застосовується в реалізаціях Anomaly Detection як один із етапів попередньої обробки, де підмножина предиктора (незалежні атрибути) знаходить шляхом видалення предикторів з нижчою або нерелевантною інформацією. Однак дуже мало технік прогнозування виконують Anomaly deatection перед тим, як тренувати моделі прогнозування.

3.2 Методи пошуку Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) та Genetic Algorithm (GA)

Декілька евристичних підходів до оптимізації були ефективно впроваджені як методи пошуку великих данніх. Наприклад, ці методи містять Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) та Genetic Algorithm (GA)

GA отримав велику увагу завдяки своїй працездатності та надійній пошуковій здатності. GA є одним із імовірнісних алгоритмів пошуку штучного інтелекту і широко застосовується для ряду задач оптимізації.

Він був натхненний виживанням найбільш підходящого принципу еволюційної теорії Чарльза Дарвіна та генетики.

Пошук GA починається з випадково вибраного набору осіб, який називається початковою сукупністю. Потім він повторює пошук найкращого індивіда (рішення) за допомогою трьох основних послідовних операторів - відбору, кросоверу та мутації. GA використовує індекс продуктивності, який називається функцією фітнесу, для обчислення фізичної готовності людей за ітерації. BGA - це спеціальна версія GA, яка працює, представляючи спочатку заданий простір ознак (хромосоми або розчини-кандидати) у двійкових бітових рядках. Це робить BGA добре придатним для проблем великих данних, ніж звичайний GA.

У цій роботі пропонується метод вибору гібридних функцій на основі машинного навчання для отримання найбільш релевантних та непотрібних функцій для вдосконаленого короткострокового прогнозування потреби в електроенергії в децентралізованих енергетичних системах. У запропонованому методі для процесу вибору ознак застосовується бінарний генетичний алгоритм (BGA), а для вимірювання оцінки придатності ознак використовується регресія процесу Гауса (GPR).

Наскільки ми розуміємо, існує дуже мало дослідницьких робіт, які виконували роботу з вибору особливостей до встановлення або навчання моделей прогнозування. Більше того, наскільки ми досліджували, підхід гібридного машинного навчання, заснований на BGA-GPR, ніколи не застосовувався до проблеми вибору функцій в області прогнозування попиту на електроенергію.

У цьому дослідженні залишок моделі GPR обраний як міра оцінки придатності. GPR - це потужний алгоритм регресії. Він обраний завдяки своїй вищій здатності встановлювати нелінійні співвідношення вхід-вихід на основі імовірнісних розподілів за функціями. У нього мало параметрів, які можна налаштувати та легко реалізувати. Він може дати послідовну

оцінку своєї невизначеності. GPR може безпосередньо сприймати невизначеність моделі (вхід-вихід або взаємозв'язок об'єкт-ціль).

Наприклад, він безпосередньо забезпечує розподіл для значень придатності вибору функції (помилки), а не лише одне значення як оцінку. Більшість інших інструментів ML або AI не сприймає цієї невизначеності. Більше того, GPR може додати попередні знання та специфікацію щодо поведінки взаємозв'язку вхід-вихід (об'єкт-ціль) за допомогою різних функцій ядра.

Як правило, вклади в роботі можна розглядати як (1) моделювання, параметризацію та реалізацію алгоритмів BGA та GPR відповідно до заданої проблеми вибору ознак, та (2) встановлення безшовної комбінації двох алгоритмів для єдиної роботи при вирішенні проблема вибору функції.

Зокрема, ця робота, як правило, має такі основні внески:

1. Дослідити та запропонувати доречність ефективного підходу обробки великих данних для прогнозування попиту на електроенергію з підвищеною ефективністю;
2. Покращити точність прогнозування попиту на електроенергію завдяки застосуванню Anomaly Detection перед встановленням моделей прогнозування.

3.3 метод обробки Feature Selection

Методи FS класифікуються як фільтрувальні, обгорткові та вбудовані методи [1]. Методи фільтрування не залежать від будь-якої моделі прогнозування, і вони сортують ознаки залежно від статистичних характеристик. Вони використовують оцінку кореляції для оцінки підмножини ознак. Методи FS на основі техніки фільтрування, як правило, швидкі. Підхід Filter FS містить кореляційні, взаємні інформаційні та основні аналітичні методи. Як правило, фільтри вимагають менше часу

обчислення, ніж інші техніки FS, але вони генерують набір функцій, який не підходить для конкретної моделі прогнозу.

Методи фільтрування широко застосовуються в аналізі великих даних завдяки їх обчислювальній ефективності. Методи обгортання оцінюють підмножини предикторів на основі їх вартості для конкретного прогнозуючого або класифікатора. Техніка обгортання припускає FS як пошукову задачу, яка готує різні суміші предикторів, оцінюються та протиставляються іншим сумішам.

Поширені евристичні методи оптимізації на основі штучного інтелекту, використовуються для моніторингу процедури пошуку. Порівняно з техніками фільтрування, технології обгортання виявляють поліпшену ефективність, оскільки різні набори предикторів оцінюються за допомогою прогностичної моделі або методу підгонки на кожній ітерації.

Вбудовані методи поєднують процес вибору ознак із навчанням моделі прогнозування. Наприклад, підходи до регуляризації у навчанні моделей [1] є одним із прикладів методу FS вбудованого типу. Модель LASSO, яка нормалізує параметри лінійних моделей із штрафами L1 для зменшення некорельованих коефіцієнтів до нуля, також може бути одним із прикладів вбудованого методу.

Розробка відповідного заходу оцінки фізичної форми є дуже важливою у підходах ФС. Міра оцінки придатності використовується як показник ефективності для оцінки придатності ознак кандидата. Прогнози класифікуються та відбираються відповідно до оцінених значень функції чи міри фізичної підготовки.

Комбінація предикторів, яка дає найкраще значення фітнес-міри, вибирається в кінці запущеного алгоритму FS. У цій роботі використано залишкову (похибку) моделі регресії процесу Гауса (GPR) як функцію придатності BGA. Модель GPR - це імовірнісний розподіл функцій на основі ядра. Він обраний у цій роботі завдяки своїй вищій здатності

встановлювати нелінійні відносини вхід-вихід на основі припущень імовірнісного розподілу даних вводу-виводу даних або функції. Крім того, у нього мало параметрів, які можна налаштувати та легко реалізувати.

Після всебічної оцінки вищезазначених методів ФС, заснованих на генетичному алгоритмі, ми виявляємо, що у більшості досліджень використовується звичайний генетичний алгоритм із звичайною структурою (звичайна конфігурація GA) [15]. Наприклад, початкова популяція (початковий набір хромосом) довільно створюється там, де різноманітність популяції не може бути гарантована, а поява дубльованих предикторів може вплинути на якість процедури пошуку. Більше того, звичайний GA працює з самими безперервними функціями, щоб мінімізувати бажану функцію придатності (показник оцінки FS). Це знижує ефективність алгоритму та спричинює складність обчислень та збільшує загальний час обчислень.

Припускаючи, що інтелектуальний евристичний алгоритм повинен бути найкращим варіантом для визначення цілі пошуку; існує проблема дослідження, і її слід вирішити, замінивши звичайний GA на BGA та гібридизуючи його з надійною оцінкою придатності (GPR у цьому документі). BGA спочатку представляє функції у вигляді закодованого двійкового рядка і працює з двійковими рядками, щоб мінімізувати оціночну міру на основі GPR, щоб отримати найбільш релевантну та непотрібну підмножину предикторів в кінці. BGA є більш ефективним і стабільним, ніж звичайний GA. Це також зменшує обчислювальну складність та час виконання порівняно зі звичайними GA.

3.4 Пропонований вибір функцій на основі BGA-GPR

Як показано на блок-схемі на малюнку 2, у BGA існує п’ять ключових підоперацій, а саме - кодування хромосом, обчислення об’єктивних значень, методи відбору, генетичні оператори та умова

зупинки. BGA працює над бінарним доменом пошуку (бітові струни хромосом). GA оперує кінцевим бінарним набором хромосом, заснованим на виживанні найбільш пристосованого принципу еволюційної теорії. Початкова сукупність формується та оцінюється за допомогою цільової функції. Для двійкової хромосоми, використаної в цій роботі, значення гена «1» вказує на те, що обрано специфічну особливість, на яку вказує місце «1». В іншому випадку (якщо «0»), функція не вибирається для оцінки фізичної форми.

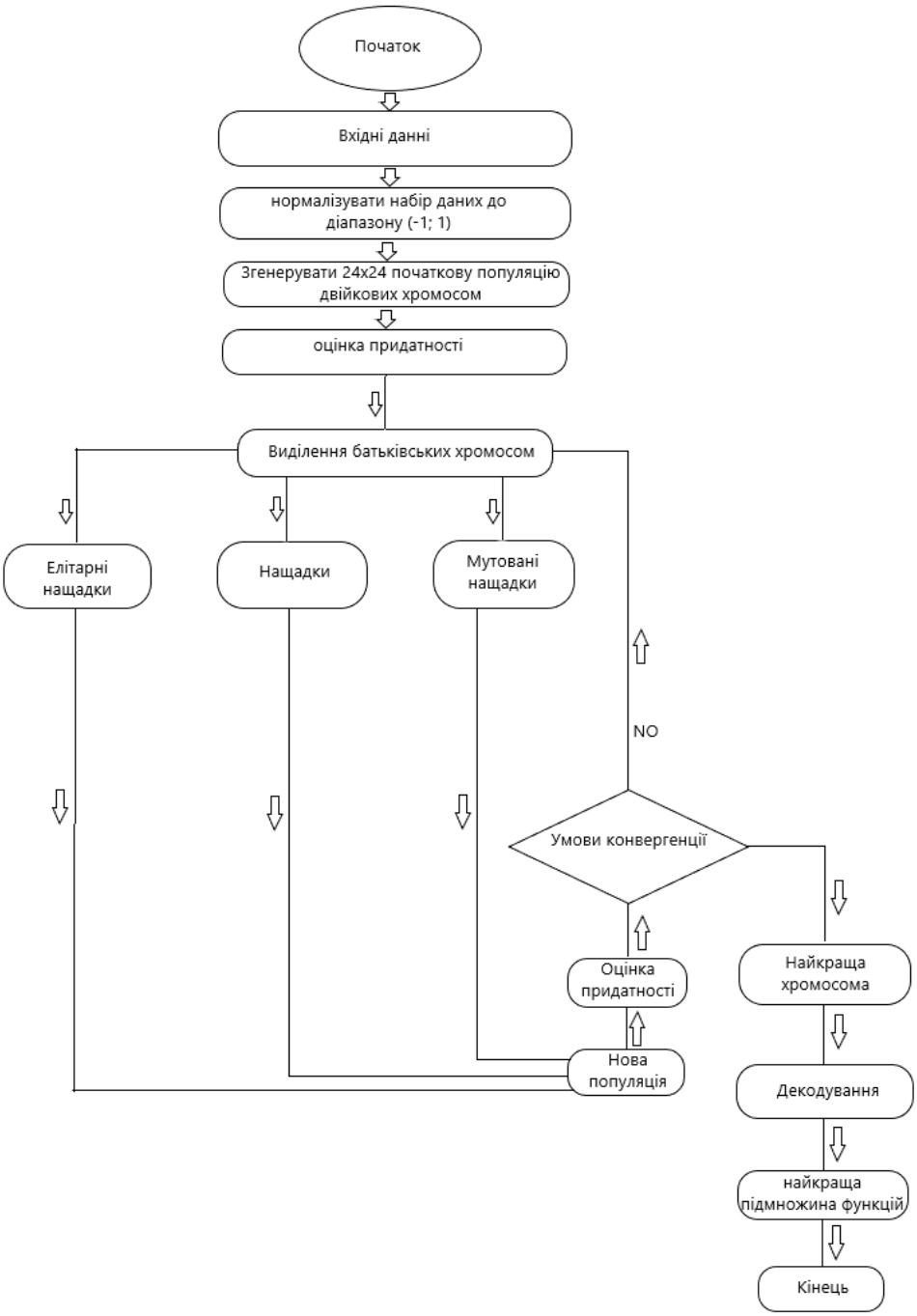

Рис. 3.1 Блок-схема схеми FS на основі BGA-GPR

Використовуючи покажчик місця змінних, вказаних «1s», особини потім упорядковуються, і відповідно до наказів вибираються верхні k найбільш придатні нащадки (елітарність розміру k), які зберігатимуться з наступним поколінням. Як тільки вибрані потомства будуть переміщені безпосередньо до наступного покоління, іншим нащадкам у поточному просторі рішень дозволено генетично переміщатися через оператори

кросовера та мутації, щоб створити потомство кросовера та мутації відповідно [26]. Потім три нащадки, а саме відбір, кросинговер та мутація, створюють новий простір рішення (нове покоління). Оператор кросинговеру - це злиття двох хромосом для створення кросоверних нащадків. Поки оператор мутації застосовується для генетичного розладу (різноманітності) генів у хромосомах шляхом підкидання бітів на основі ймовірності мутації. Дотримуючись процедур, викладених на малюнку 2, детальні механізми роботи запропонованого GA-GPR FS описані в наступних підрозділах.

Початкове населення. Простір вихідного розчину GA, що використовується в цій роботі, є матрицею розміром p × q, де p - кількість хромосом, а q - довжина хромосоми (так звана довжина генома). p дорівнює чисельності популяції, а q дорівнює кількості бітів або генів у кожної особини. Рекомендується дозволити кількість хромосом дорівнювати принаймні довжині хромосом, щоб хромосоми в кожній популяції охоплювали область пошуку [29].

Оцінка придатності. Для того, щоб BGA вибрав підмножину предиктора, слід вказати цільову функцію (драйвер BGA) для обчислення дискримінаційної сили кожного підмножини предиктора. Придатність кожної хромосоми в популяції оцінюється з використанням функції фітнесу на основі GPR. У цій роботі придатність різних підмножин функцій оцінюється за допомогою MSE (середня квадратична помилка) прогнозних залишків моделі GPR. Модель GPR f (x) підходить для кожного підмножини функцій. Отже, MSE цілі навчання та оцінка моделі GPR, оцінена для кожного підмножини ознак у просторі пошуку ознак, визначеному в Таблиці 2, використовується як міра оцінки придатності, і вона визначається наступним чином.

$$fit = \frac{1}{n} \sum_{i=1}^{n} (T_i - f_i)^2 \quad (3.1)$$

де T - вектор навчальної цілі (потреба в електроенергії), а n - кількість навчальних зразків або спостережень.

Метою BGA є мінімізація функції придатності (MSE), визначеної у рівнянні (3.1), шляхом вибору підмножини вхідних ознак, що мають найкращу придатність, у наступних ітераціях. У кожній хромосомі значення гена «1» показує конкретну змінну, яку вказує місце «1». Якщо воно дорівнює «0», провісник не вибирається для оцінки відповідної хромосоми. Хромосоми, що представляють провісники, кодуються як бітові струни. У міру запуску BGA оцінюються окремі хромосоми (підмножини ознак) у поточній популяції, і їх придатність оцінюється на основі залишків або помилок моделі GPR. Хромосоми з меншою придатністю (менші залишкові або помилкові) мають більшу ймовірність збереження з наступною популяцією або пулом спарювання.

Кожна ітерація запущеного BGA гарантує, що BGA зменшує рівень помилки і елітує хромосому з найнижчим (найкращим) значенням цільової функції. Індивідуальна хромосома, що відповідає найменшому рівню похибки оцінки придатності, містить бажані найбільш відповідні ознаки.

Висновки

У цьому розділі обговорюється тематичне дослідження запропонованої роботи. У цьому розділі також представлені порівняльна перевірка, оцінка результатів FS для вдосконаленого прогнозування та кількісний аналіз релевантності результатів FS. Механізм відбору повинен бути заздалегідь визначений у BGA, щоб забезпечити безперервне покращення популяції протягом усіх показників фізичної підготовки або ітерацій. Метод відбору допомагає BGA ігнорувати найгірших особин і зберігати лише найкращі хромосоми. Існує декілька методів відбору для BGA, однак у цьому дослідженні використовується Механізм відбору через його простоту використання, швидкість та ефективність. Крім того, метод

відбору накладає на BGA краще обтяжувальне відбір, що призводить до кращого рівня збіжності та гарантує, що рішення з поганими кандидатами не переходять у наступне покоління. У турнірному відборі розміру 2 обираються дві особи з простору рішень після виведення елітних нащадків.

РОЗДІЛ 4

ОБРОБКА ДАНИХ МЕТОДАМИ ANOMALY DETECTION

4.1 Детектування аномалій

Для пошуку аномалій в машинному навчанні існує два методи аналізу даних: детектування аномалій (Outlier Detection) і детектування новизни (Novelty Detection).

В першому випадку детектується вихід значень досліджуваних параметрів за межі області, яка в результаті аналізу була встановлена як «нормальна». В другому випадку реєструється поява «нового об'єкту» - не аномального об'єкту відсутнього у базі даних.

Як «викид», так і «новий об'єкт» відрізняється від об'єктів наявної навчальної вибірки. Проте на відміну від викиду, у майбутньому цей об'єкт повинен бути включеним до навчальної вибірки. Таким чином, межі «області нормальності» динамічно змінюються, вміщуючи виявлені нові об'єкти.

Наприклад, при аналізі виміряних значень температури навколишнього середовища відкидаються аномально великі або маленькі значення, що відноситься до детектування аномалій. Якщо ж алгоритм аналізу передбачає оцінку кожного нового значення у термінах «схожості» на наявні значення у навчальній вибірці – це детектування новизни.

В термінах задачі виявлення аномальної важливим є детектування як аномалій значень на координатній площині параметрів, так і детектування новизни – поява нових поведінкових шаблонів, що є відмінними від наявних у навчальній вибірці та підлягають обробці як новий варіант «норми».

Викиди виникають внаслідок:

- помилок у даних (неточності вимірювання, округлення, невірного запису);

- наявності завад, що спричиняють невірну класифікацію об'єктів, тобто їх помилкове віднесення до певних груп;

- присутності об'єктів «сторонніх» вибірок (наприклад, даних з датчика, що вийшов з ладу).

Суб’єктивні фактори, в свою чергу, здатні обумовити появу аномалій внаслідок помилкової інтерпретації або несанкціонованого втручання людини у роботу технічного обладнання.

При аналізі поведінкових характеристик постає задача вибору з усієї сукупності параметрів, що описують поведінку та переміщення людини, саме тих вирішальних параметрів, на підставі яких можна детектувати викиди або новизну.

В найпростішому випадку одним з таких вирішальних параметрів може виступати кількість спрацювань датчика руху в одному і тому самому приміщенні протягом певного невеликого інтервалу часу τ (наприклад, τ=10 хв), а другим – усереднене значення енергії споживання за цей самий інтервал. Такий підхід дозволяє виявити випадки швидкого «без системного» переміщення, що може бути наслідком знаходження у стані афекту, наляканості чи психологічного розладу, адже однакова кількість спрацювань одного датчику протягом 10 хвилин розглядається як «аномалія», а протягом години – як «норма».

Для моделювання найпростішого варіанту аналізу за допомогою методу Anomaly Detection визначаємо такі дані: Ознака 1 - кількість N спрацьовувань датчика руху протягом кожного з послідовних періодів тривалістю 5 хвилин, Ознака 2 – середнє значення електроспоживання W за обраний нами період. Значення цих ознак за період спостереження утворюють зразкову матрицю – множину двоелементних векторів (K=2 – кількість ознак, що беруться до розгляду). Приклад накопичених даних значень обраних нами ознак наведено у табл. 3.1.

Таблиця 3.1

Приклад накопичених даних

№	Час	W, Вт	Кількість, N
1	00:00-00:05	20	50
2	00:05-00:10	30	40
3	00:10-00:15	25	35
4	00:15-00:20	27	40
...	...	...	...
127	10:30-10:35	20	10
128	10:35-10:40	25	30
...	...	...	...

Отримані результати наносяться на координатну площину параметрів N та W. В подальшому графічно-аналітичними методами Anomaly Detection визначається «зона нормальності» (рис.3.1).

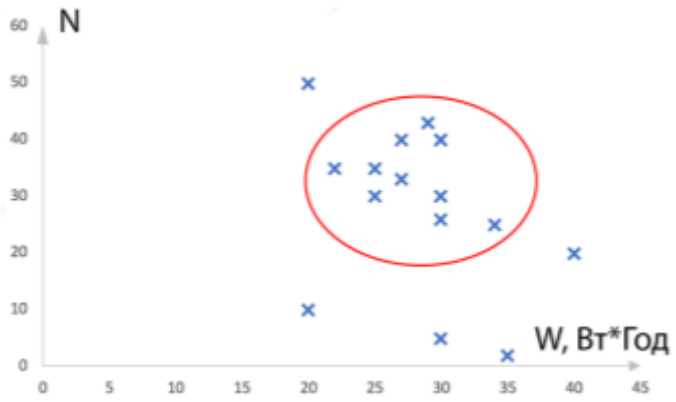

Рис. 3.1 Приклад зони нормальності

Для формування цієї зони використовуються різні підходи, зокрема, з урахуванням експертних оцінок та даних попередніх періодів

спостережень. Наявні на графіку відхилення вважаються аномаліями і потребують втручання інших сервісів інформаційно-програмного забезпечення MicroGrid для подальшого відпрацювання – активації тривоги, сповіщення користувача, тощо.

Виявлення меж «зони нормальності» може бути полегшено за рахунок попередньої кластеризації. Тоді кластери суттєво меншого розміру, швидше за все, відносяться до аномалій.

Для навчання алгоритмів машинного навчання величезну кількість даних збирається з багатьох датчиків в IoT. Однак через використання цього великої кількості даних кращий алгоритм навчання контролюється неконтрольованим алгоритмом навчання. Дійсно, алгоритм кластеризації має можливість ефективно обчислювати дані і групувати подібні шаблони активності користувачів в кластери.

Використання алгоритму кластеризації повинно призвести до виявлення тимчасових відносин шляхом обробки сприйманих даних. Дійсно, обробка сприйманих даних являє собою перший крок у визначенні активності користувачів в розумному будинку. Таким чином, алгоритм кластеризації моделі відповідає методології, яка складається з етапу перетворення сприйманих даних датчиків, найбільш частому спостереженні за схемою і етап видобутку та аналогічний етап групування.

У реальній системі MicroGrid кількість точок спостереження, що складають навчальну вибірку, і відповідно, кількість рядків у табл.3.1 буде збільшуватися. Постійне накопичення даних дозволяє збільшувати обсяг навчальної вибірки та відповідно динамічно змінювати «зону нормальності».

4.2 Метод обробки даних Anomaly Detection

Метод обробки у масиві даних базується на емпіричному правилі «трьох сигм», або «68-95-99,7»

У статистиці правило «68-95-99,7» - це скорочення, яке використовується для вказання ймовірності подій, які характеризують знаходження аналізованих значень в межах смуги навколо середнього значення при нормальному законі розподілу випадкової величини з шириною двох, чотирьох та шести стандартних відхилень, відповідно. Точніше, ці імовірності складають 68,27%, 95,45% та 99,73%, тобто відповідні частки всіх значень випадкової величини лежать в межах одного, двох та трьох стандартних відхилень від середнього значення відповідно.

При проведенні обчислень за одиницю виміру відхилення випадкової величини, що підпорядковується нормальному закону розподілу, від центру величини розсіювання (математичного очікування) приймається середньоквадратичне відхилення σ . Тоді на підставі виразу:

$$P(-1 < \bar{x} < l) = \widehat{\Phi}(\frac{l}{E}) \qquad (4.1)$$

випливають корисні при різних обчисленнях рівності:

$$P(-\sigma < \bar{x} < \sigma) = \Phi(\frac{1}{\sqrt{2}})=0.683,$$

$$P(-2\sigma < \bar{x} < 2\sigma) = \Phi(\sqrt{2})=0.954,$$

$$P(-3\sigma < \bar{x} < 3\sigma) = \Phi(\frac{3}{\sqrt{2}})=0.997,$$

Ці результати геометрично зображені на рис 4.1.

Майже достовірно, що випадкова величина (помилка) не відхиляється від математичного очікування по абсолютній величині більше ніж на 3σ . Це припущення називається “правилом 3х сигм”.

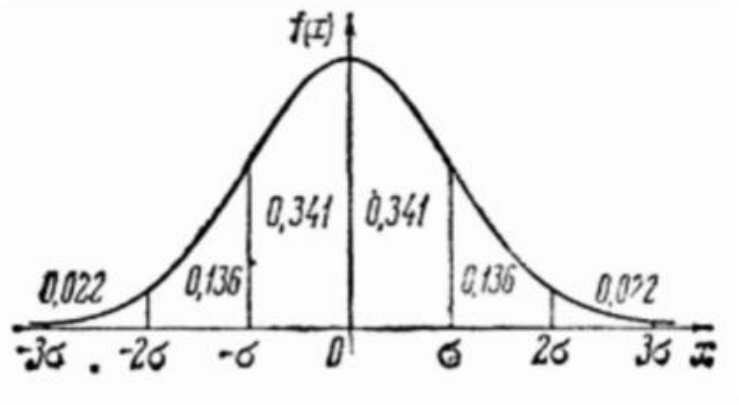

Рис. 4.2 Нормальний розподіл випадкової величини.

Для приблизної нормальної області даних значення в межах одного стандартного відхилення середнього значення складають близько 68% встановленого; в межах двох стандартних відхилень - близько 95%; в межах трьох стандартних відхилень - близько 99,7%. Наведені значення являють собою теоретичні ймовірності, призначені для наближеної оцінки емпіричних даних, отриманих від нормальної вибірки.

Метод обробки на базі правила трьох сигм. Обробка даних базується на правилі 3-х сигм. Для правильної обробки за допомогою правила трьох сигм потрібно порівнювати тренувальні дані та дані які надходять в реальній момент часу, для визначення аномальних результатів.

Для обробки даних потрібно врахувати математичне очікування:

$$M_x = \xi = M_1 = \sum_i xp(x) = \frac{1}{m}\sum_{i=1}^{n} x_i\,;$$

Наступним етапом є визначення середньоквадратичного відхилення, для цього потрібно визначити математичне очікування в квадраті M_x^2та дисперсію D_x.

$M_x^2 = M_2 = \frac{1}{m}\sum_{i=1}^{n} x_i^2$; M_2- математичне очікування в квадраті;

$$D_x = \sigma^2 = M_x^2 - \xi^2 = M_2 - M_1^2;$$

Після врахування математичного очікування та дисперсії можемо вивести формулу середньоквадратичного відхилення σ:

$$\sigma = \sqrt{M_2 - M_1^2};$$

Отже, спираючись на правило трьох сигм ми можемо описати основний критерій обробки аномальних даних:

$$value <= 3\sigma + M_1$$

value - це значення яке надходить на перевірку.

σ- середньоквадратичне відхилення,

M_1-математичне очікування.

Цей вираз дає нам змогу виявити дані які будуть аномальними, якщо значення які надійшли на перевірку виходять за межі тестового значення $3\sigma + M_1$, ці дані являють собою аномалію.

Такий підхід обробки даних можна реалізувати на мовах програмування з наявними бібліотеками математичних операцій.

Головним критерієм для використання цього методу є підготовка тренувальних даних, які мають бути зібрані впродовж використання будь якими системами MicroGrid, де можуть бути наявні аномальні дані.

4.3 Приклад сервісу обробки даних

Сервіс обробки великих даних можливо розгорнути на будь яких системах MicroGrid в яких присутні датчики які мають функції відстеження активності. Функціонування такого сервісу машинного навчання та його взаємодія з MicroGrid наведено на рис. 4.3

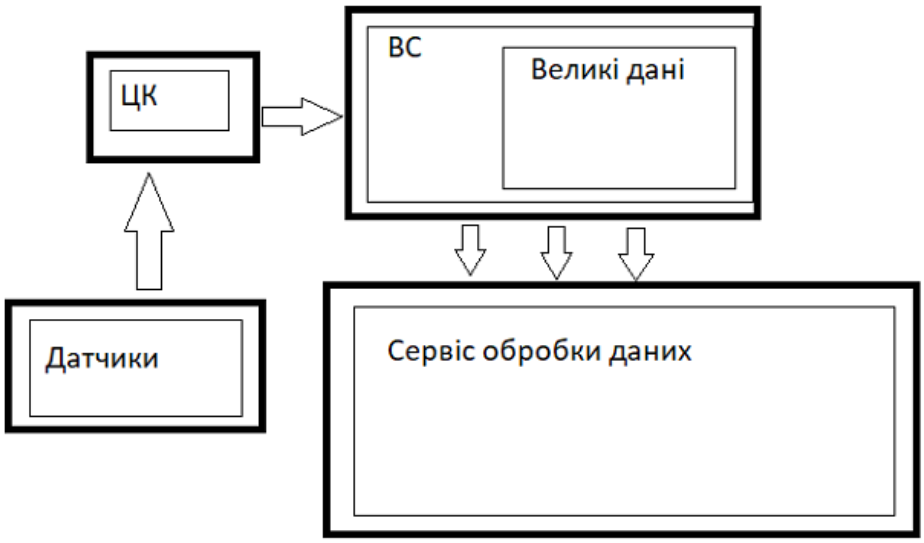

Рис. 4.3 Робота сервісів машинного навчання у MicroGrid

Датчики надсилають дані у центральний контроллер, необхідно зв’язати датчики та ЦК за допомогою бездротової передачі даних локальної мережі (Wi-Fi). Для цього до кожного датчику необхідно під’єднати модуль, що вміє відправляти і отримувати інформацію в локальну мережу або в інтернет за допомогою Wi-Fi. У свою чергу, центральний контроллер, підключений до мережі інтернет, збирає дані з датчика і надсилає їх на віддалений сервер (ВС). Усі дані, що фіксуються з центрального контроллера, надходять на ВС у реальному часі. ВС через спеціальний ключ доступу додає дані у не реляційну базу даних, що з часом стає дуже великою через великий обсяг даних і відноситься до категорії BigData[12].

Сервіс, зображений на рис. 3.3, може знаходитись на віддаленому сервері, іншому сервері, або на персональному комп’ютері користувача. Сервіс являє собою Back-End частину, де здійснюється генерація часових інтервалів за кожен день, де система обробляє дані отримані з бази даних віддаленого сервера і записує, скільки зафіксовано спрацювань датчиків за n-інтервал часу, при цьому фіксується дата та день тижня. Наступним кроком є створення векторів інтервалів часу різних днів, тобто створюється

окремий вектор для кожного проміжку часу на основі обробленої бази з попереднього кроку, в які заносяться кількість спрацювань датчиків за даний проміжок, за всі дні. Далі система детектує аномалії за певні проміжки часу: оброблює кожен проміжок часу за певний день тижня, з цього вектору формує тренувальний вектор . Дані тренувального вектору порівнюються з даними щодо кількості спрацювань за певний проміжок часу поточної ітерації перебору об’єктів даних дня. Якщо аномалію зафіксовано, поточний вектор заноситься у базу аномалій.

Машинне навчання в сервісі оборки даних. Логічну схему послідовності дій машинного навчання зображено на рис.3.4

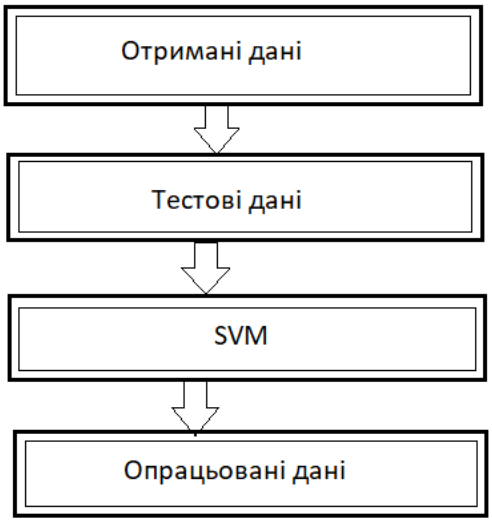

Рис. 3.4 Машинне навчання

1. Блок “Отримані дані” – накопичені дані після вимірювання (Big Data).
2. Блок “Тестові дані” – дані що надходять до системи для виявлення: чи є аномалія у даному наборі даних.
3. Блок “SVM” (Support Vector Machine) контролюється заданими моделями та алгоритмами навчання, які аналізують тренувальні та тестові дані, що використовуються для класифікації та регресійного аналізу. Враховуючи набір навчальних прикладів, кожен з яких позначається як такий, що

належить до однієї з двох категорій, SVM-модель являє собою представлення тестового прикладу як точки в просторі, мапованому таким чином, що приклади окремих категорій діляться явним розривом. Нові приклади потім вносяться у той самий простір і беруть участь у подальшому аналізі вже як елементи навчальної вибірки.

4. Блок “Отримані дані” – оброблені дані, виявлені аномалії після класифікації даних SVM.

В роботі запропоновано застосування Anomaly Detection як один із засобів Machine Learning для обробки даних (BigData) у MicroGrid. Наукова новизна отриманих результатів полягає у отриманні подальшого розвитку теорії застосування методів машинного навчання для обробки та аналізу великих даних (Big Data) у MicroGrid.

Запропоновано розв’язання задачі пошуку аномалій в роботі датчиків на базі застосування методу опорних векторів (SVM). Сформований метод має великий потенціал для подальшого розвитку і використанні його в різних сферах MicroGrid, для покращення обробки великих даних та детектування аномалій. Метод дає можливість виявляти аномалії в роботі датчиків та сприяти покращенню енергоефективності систем та подальшого спостереження за системами в автоматичному режимі без втручання людини.

Даний метод відкриває можливості для розробників програмного забезпечення, для подальшого його розвитку та вдосконалення, та застосування з будь-якими програмними засобами та мовами програмування.

Можливості методу універсальні та підійдуть як для простих систем, наприклад розумних будинків, так і для великих підприємствам з великим обсягом датчиків і систем керування MicroGrid. Саме великим системам буде доцільно використовувати методи обробки даних з пошуком

аномалій, для їх швидкого детектування та усунення з метою збереження витрат на енергію, та покращення енергоефективності.

4.4. Модифікований метод обробки в MicroGrid.

Модифікація методу обробки даних. Огляд методів SVM та FS показав що рішення використовувати їх в парі як гібридний метод є доцільним, тому запропоновано рішення для реалізації сервісу обробки великих данних для прогнозування навантеження на основі машинного навчяння для систем MicroGrid.

Отже опишемо нову схему функціонування такого сервісу машинного навчання та його взаємодію з MicroGrid зображену на рис.3.5.

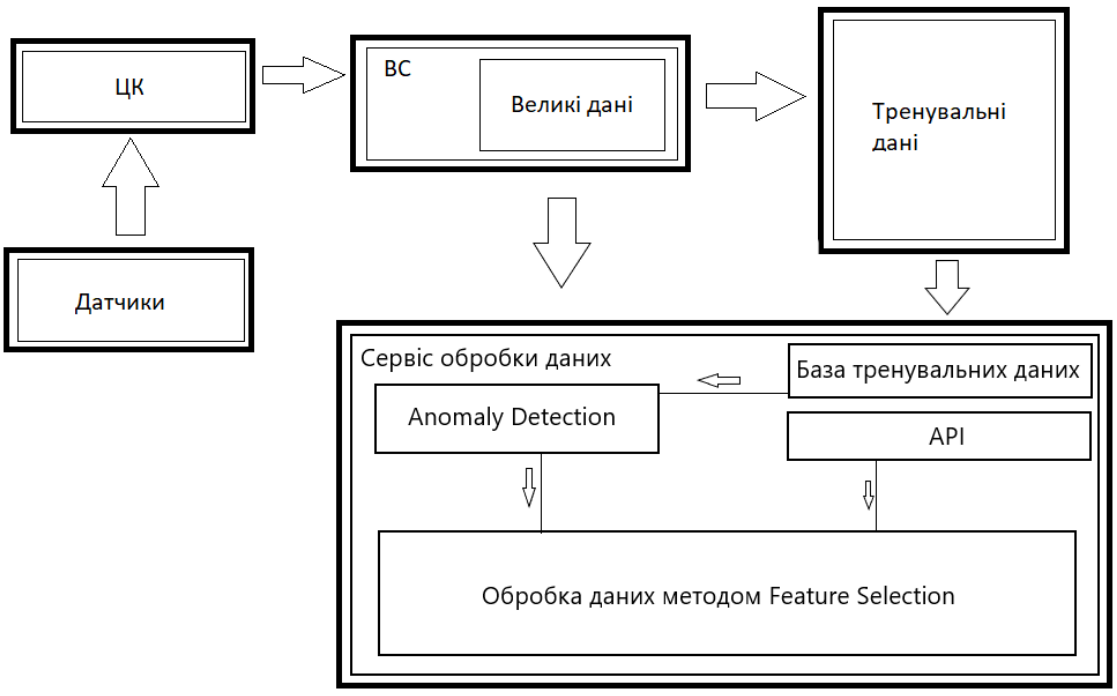

Рис 3.5 Функціонування сервісу з блоком обробки даних

Функціонування сервісу має новий підхід і додає нові блоки для покращення обробки даних. Який під час роботи сервісу буде отримувати з віддаленого серверу дані з ких буде робити вибірку тренувальних даних та відправляти їх в сам сервіс в якому наявна база тренувальних даних, а

також буде присутній блок API щоб забеспечити систему необхідними статистичними даними. Цей блок збирає і оновлює як і звичайні дані так і аномальні результати, ти самим оновлюючи і виключаючи недолік методу SVM, а саме необхідність постійно оновлючати базу тренувальних даних.

Також представимо оновлену схему машинного навчання на рис 3.6.

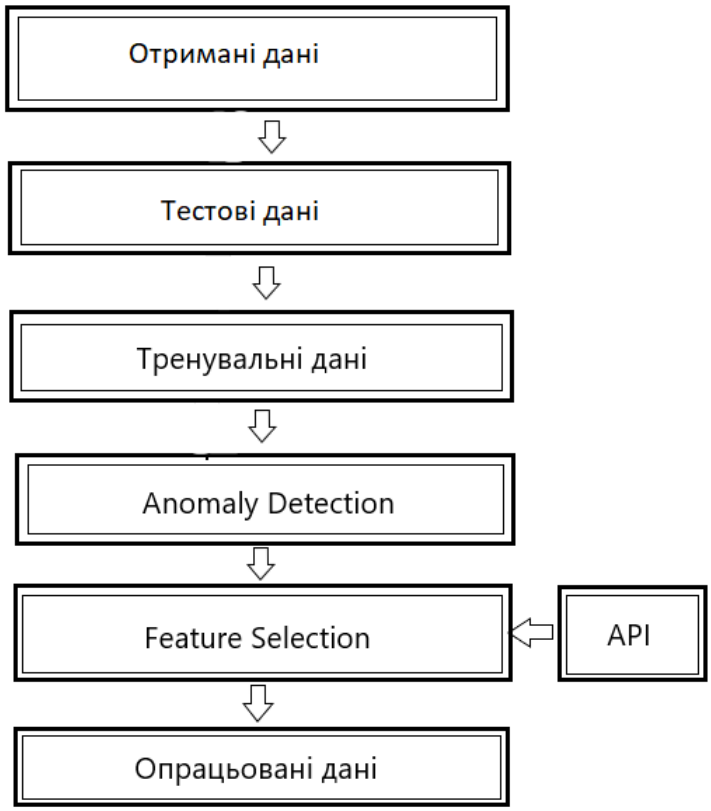

Рис 3.6. Машинне навчання з блоком тренувальні дані.

- Блок “Отримані дані” – накопичені дані після вимірювання (Big Data).
- Блок “Тестові дані” – дані що надходять до системи для виявлення: чи є аномалія у даному наборі даних.

- Блок “Тренувальні дані” – дані накопичені за час роботи сервісу.
- Блок “Anomaly Detection” (Support Vector Machine) контролюється заданими моделями та алгоритмами навчання, які аналізують тренувальні та тестові дані, що використовуються для класифікації та регресійного аналізу. Враховуючи набір навчальних прикладів, кожен з яких позначається як такий, що належить до однієї з двох категорій, SVM-модель являє собою представлення тестового прикладу як точки в просторі, мапованому таким чином, що приклади окремих категорій діляться явним розривом. Нові приклади потім вносяться у той самий простір і беруть участь у подальшому аналізі вже як елементи навчальної вибірки.
- Блок “Feature Selection” – обробляї зібрані дані системи оброблені методом SVM, отримує дані з яких уже відкинуті аномальні дані через що може робити більш точні висновки щодо навантаження системи, також отримує необхідні статичні та статистичні данні з API які потрібні для точного прогнозувння навантаженя системи.
- Блок “Отримані дані” – оброблені дані, данні прогнозованого навантеження системи з відкинутіми аномаліями, дані можуть використовуватися для розрахунку затрат енергії на роботу системи.

Зі схеми видно, що тепер при обробці даних сервіс також використовує накопичені дані, тим самим зменшує ймовірність появи помилок при прогнозуванні навантаження в системі MicroGrid.

Висновки

Описано алгоритм детектування аномалій, змодельовано варіант обробки даних методом Anomaly Detection та Feature Selection. Описано метод пошуку аномалій за допомогою правила трьох сигм, суть якого полягає в порівнянні тренувальних даних та даних які надходять в реальній момент часу, для визначення аномальних результатів. Наведено приклад сервісу обробки даних для прогнозування навантаження з описом машинного навчання. Описано метод покращення

сервісу обробки даних включення в нього блоку обробки та оновлення бази тренувальних даних, також доданий блок API для більш точного прогнозування навантаження.

РОЗДІЛ 5

РОЗРОБКА СТАРТАП – ПРОЕКТУ

Однією з головних причин створення, успішного розвитку та подальшого існування стартапів вважається нечесність і повільність основних корпорацій, які успішно використовують існуючі продукти, а розробка та створення нових майже не бере участі. Тому в плані впровадження нових ідей їх мобільність викликає конкуренцію з великими корпораціями.

Основним джерелом для початку є гарна винахідлива ідея. Насправді більшість людей дотримуються свіжих і незвичних ідей і часто купують їх, не економлячи великі суми. Багато ідей, які не мають фізичного втілення, а існують лише на папері або "словами" (планами запуску), можуть коштувати багато. Іншим аспектом успіху цієї ідеї є її попит (ступінь, необхідний для споживача), оскільки ідея може бути незвичною та новою, але користь буде невеликою.

Етапи розроблення стартап-проекту:

1. Маркетинговий аналіз стартап-проекту

1.1 Розробляються деталі проектної ідеї та визначаються загальні напрямки використання потенційних товарів чи послуг, а також їх відмінності від конкурентів;

2.1 аналізуються ринкові можливості щодо його реалізації;

3.1 На основі аналізу ринкового середовища розробляються стратегії впровадження на ринок потенційних товарів у рамках проекту.

2. Організація стартап-проекту

2.1 складається календарний план-графік реалізації стартап-проекту;

2.2 Розраховується потреба в основних фондах та нематеріальних активах;

2.3 Визначається планова кількість виробництва потенційних товарів, на основі якої визначається потреба у фізичних ресурсах та персоналі;

2.4 Розраховуються загальні початкові витрати на початок проекту та заплановані загальні витрати, необхідні для реалізації плану.

3. Фінансово-економічний аналіз та оцінка ризиків проекту

3.1 визначається обсяг інвестиційних витрат;

3.2 Розраховуються основні фінансово-економічні показники проекту (обсяг виробництва, собівартість продукції, ціна реалізації, податкове навантаження та чистий прибуток) та визначаються показники інвестиційної привабливості проекту (запас фінансової стійкості, прибуток від продажу та інвестицій, термін окупності проекту);

3.3 Визначено рівень проектного ризику, визначено основні ризики проекту та заходи їх попередження (реагування на ризик).

4. Заходи з комерціалізації проекту

4.1 визначення цільової групи інвесторів та опису їх ділових інтересів;

4.2 складання інвест-пропозиції (оферти): стислої характеристики проекту для попереднього ознайомлення інвестора із проектом;

4.3 планування заходів з просування оферти: визначення комунікаційних каналів та площадок та планування системи заходів з просування в межах обраних каналів;

4.4 Ресурсний план реалізації заходів щодо просування пропозиції. Ці кроки, що здійснюються постійно та вчасно, створюють передумови для успішного запуску ринку. Перший етап розробки стартап-проекту буде розглядатися в рамках магістерської дисертації, тобто аналізу ринку стартового проекту системи для виявлення розбіжностей у поведінці особистості.

5.1. Опис ідеї проекту (технології)

Спочатку було розроблено зміст ідеї та можливі базові потенційни ринки даного проекту в межах яких слід шукати потенційних клієнтів. В табл. 5.1 наведено напрямки застосування та вигоди для користувача.

Таблиця 5.1

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Зміст ідеї: створення та реалізація алгоритму прогнозування навантаження системи за допомогою виявлення аномалій в BigData за допомогою методів Machine Learning;	Децентралізовані системи MicroGrid.	Прогнозування навантаження системи.

Було визначено, що ідея має здатність конкурувати та має унікальні для ринку якості.

5.2. Технологічний аудит ідеї проекту

Наступним кроком було розроблено технологічний аудит ринку для визначення доступності технологій для розробки програмного забезпечення (табл. 5.2).

Таблиця 5.2

№ п/п	Ідея проекту	Технології її реалізації	Наявність технологій	Доступність технологій
1	Детектування аномалій за допомогою Machine Learning використовуючи SVM	Розробка, дослідження, програмування;	Не наявна	Доступна
2	прогнозування навантаження системи за допомогою гібридно методу Anomaly Detection та Feature Selection	Розробка, дослідження, програмування;	Не наявна	Доступна
3	Вивід результатів, оптимізація та розробка вебдодатку для відображення даних	Розробка, дослідження, програмування;	Не наявна	Доступна

Було визначено, що продукт можна впровадити технічно, але для досягнення цієї мети необхідно було розробити відповідну технологію.

5.3. Аналіз ринкових можливостей запуску стартап-проекту

Аналіз виявив можливості, які можуть бути використані на проектному ринку під час реалізації ринку, та ринкові ризики, які можуть перешкоджати реалізації проекту. Це дозволяє планувати керівні принципи розробки проектів, враховуючи кон'юнктуру ринку, потреби потенційних замовників та пропозиції конкурентних проектів (табл. 5.3)

Таблиця 5.3

№ п/ п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	300
2	Загальний обсяг продаж, грн/ум.о	13000
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу (вказати характер обмежень)	Немає
5	Специфічні вимоги до стандартизації та сертифікації	Є
6	Середня норма рентабельності в галузі (або по ринку), %	70

Тому динаміка ринку збільшується, а загальний обсяг продажів великий, тому розвиток ідеї цього проекту є перспективним та доцільним.

Далі визначаються потенційні групи споживачів, їх характеристики та формується орієнтовний перелік вимог до продукції для кожної групи (табл. 5.4).

Таблиця 5.4

Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
Спостереження за станом MicroGrid; передвчасне попередження про проблеми; прогнозування в реальному часі	Простий користувач (для домашніх умов); Компанії енергопостачання для прогнозування необхідності постачання енергії.	Експлуатація як у динаміці так і у статиці, в різних, як складних, так і у умовах спокою;	- до продукції: Точність; Надійність; Дешевизна; Якість; - до компаніїпостачальника: Точність; Брендінг та відомість; Гарантійність

Далі аналіз визначив цільову аудиторію та її потреби. Після виявлення потенційних груп споживачів аналізується ринкове середовище: таблиця факторів (таблиця. 5.5), що сприяють реалізації проекту на ринку та сукупність факторів.

Таблиця 5.5

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Перехват передачі інформації	Складність персоналізації інформації	Налагодження системи захисту інформації; розширення серверів для зберігання інформації
2	Конкуренція	Складність персоналізації інформації	Налагодження системи захисту інформації; розширення серверів для зберігання інформації
3	Зберігання інформації	Складність персоналізації інформації	Налагодження системи захисту інформації; розширення серверів для зберігання інформації

Було визначено, що загрози і можливості можливо фізично подолати. Надалі проводиться аналіз пропозиції: визначаються загальні риси конкуренції на ринку (табл. 5.6)

Таблиця 5.6

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
1. Тип конкуренції: чиста	В кого краще - в того купують	Покращення товару та сфери обслуговування
2. За рівнем конкурентної боротьби: локальна	Належить до повсякденного ринку збуту;	Розширення функціоналу та орієнтації користувачів
3. За галузевою ознакою: міжгалузева	Притаманна різним галузям застосування;	Розширення функціоналу та галузей застосування
4. Конкуренція за видами товарів: товарно-родова та товарно-видова	Належить до аналізаторів поведінки людини	Розширення фукціоналу пристрою
5. За характером конкурентних переваг: цінова та нецінова	Чим дешевше – тим привабливіше; Чим краще – тим рентабельніше;	Покращення цінової політики та якості товару
6. За інтенсивністю: не марочна	Не жорстка конкуренція	Агресивні та не агресивні форми піару

Заключним етапом ринкового аналізу можливостей реалізації проекту є складання SWOT-аналізу (табл. 9.9) - матриці сили (Strength) та аналізу слабких місць (Weak), загроз (Troubles) та можливостей (Opportunities).

Таблиця 5.9

Сильні сторони:	Слабкі сторони:
Навчання системи за допомогою Machine Learning ; Прогнозування навантаження системи на основі Feature Selection; Обробка великих данних (Big Data)	Не захищені дані; Необхідно багато серверів для зберігання Big Data
Можливості:	Загрози:
Покращення безпомилковості інформації; Надійність захищеності інформації при її передачі;	Передача інформації; Зберігання інформації; Захист інформації;

Висновки

Згідно з проведеним аналізом, у розробленому плані є можливість комерціалізації ринку. Зростаючий попит на подібні послуги додає масового придбання цього програмного забезпечення, але створює жорсткі умови конкуренції для виходу на ринок, де динаміка ринку сприяє проекту, що розробляється.

Цей проект має великий потенціал для реалізації, маючи на увазі потенційну групу клієнтів. Бар'єрами для виходу на ринок може бути відсутність великого виробника, жорсткий конкурентний тиск великих фірм подібної продукції. Але якщо ви агресивно боретеся в конкурентному середовищі, проект із сертифікації бренду має великі можливості та можливості, де ринок отримує місце в економіці в майбутньому. Подальша реалізація проекту є здійсненною та економічно ефективною.

ЗАГАЛЬНІ ВИСНОВКИ ПО РОБОТІ

В дипломній роботі описано метод обробки даних у мережі MicroGrid за допомогою методів машинного навчання.

1. Наявність аномалій у MicroGrid призводить до необхідності оперування великими обсягами даних, що обгрунтовує доцільність застосування спеціалізованих методів роботи з великими даними (Big Data).

2. Серед сукупності методів машинного навчання найбільш придатним є метод детектування аномалій (Anomaly Detection), який було обрано для визначення нетипових та аномальних результатів у MicroGrid. Огляд та порівняльний аналіз чотирьох методів машинного навчання показав, що найбільш доцільним серед них для вирішення поставленої задачі є метод Support Vector Machine (SVM).

3. Децентралізовані оператори енергетичних систем, агрегатори, постачальники, менеджери та інші зацікавлені сторони зазнають труднощів через кілька конфліктів, що варіюються від недостатнього постачання електроенергії до зростаючого споживання. Тому нещодавнє впровадження децентралізованих енергетичних систем вимагає відповідних та застосовних інструментів вибору функцій (FS) та моделей прогнозування для економічного та ефективного моделювання споживання.

4. В результаті розглядання видів аномалій у MicroGrid розроблено структуру окремих блоків сервісу машинного навчання, запропоновано метод пошуку аномалый на базі правила трьох сигм, який детектує аномальні значення роботи систем MicroGrid та гібридній метод порогнозування навантаження в системах MicroGrid. Наведено приклад сервісу обробки даних та описано алгоритм машинного навчання. Описано метод покращення сервісу обробки даних включення в нього блоку Feature Selection та оновлення бази тренувальних даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Програмне забезпечення для макромоделювання системи керування MicroGrid / Ю. С. Ямненко, А. В. Моргун, О. М. Комаревич // Electronics and communications. - 2016. - Т. 21, № 6. - С. 61-66. - Режим доступу: http://nbuv.gov.ua/UJRN/eisv_2016_21_6_10.
2. Komarevych, O. Виявлення аномальної поведінки людини у MicroGrid на базі машинного навчання / Oleksandr Komarevych, Yuriy Khokhlov, Yuliia Yamnenko // Мікросистеми, Електроніка та Акустика. – 2018. – Т. 23, N 4. - С. 36-41. – Режим доступу : DOI : 10.20535/2523- 4455.2018.23.4.143310.
3. Mozer, M.C. The neural network house: An environment hat adapts to its inhabitants. Proc. AAAI Spring Symp. Intell. Environ. 1998, 110–114.
4. Das, S.K.; Cook, D.J.; Battacharya, A.; Heierman, E.O.; Lin, T.Y. The role of prediction algorithms in the MavHome smart home architecture. IEEE Wirel. Commun. 2002, 9, 77–84.
5. Williams, G.; Doughty, K.; Bradley, D. A systems approach to achieving CarerNet-an integrated and intelligent telecare system. IEEE Trans. Inf. Technol. Biomed. 1998, 2, 1–9.
6. Korhonen, I.; Lappalainen, R.; Tuomisto, T.; Kööbi, T.; Pentikäinen, V.; Tuomisto, M.; Turjanmaa, V. TERVA: Wellness monitoring system. In Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Hong Kong, China, 1 November 1998; Volume 4, pp. 1988–1991
7. Valtonen, M.; Vuorela, T.; Kaila, L.; Vanhala, J. Capacitive indoor positioning and contact sensing for activity recognition in smart homes. J. Ambient Intell. Smart Environ. 2012, 4, 305–334.
8. Zhu, N.; Diethe, T.; Camplani, M.; Tao, L.; Burrows, A.; Twomey, N.; Kaleshi, D.; Mirmehdi, M.; Flach, P.; Craddock, I. Bridging e-health and

the internet of things: The sphere project. IEEE Intell. Syst. 2015, 30, 39–46.

9. Van Hoof, J.; Kort, H.; Rutten, P.; Duijnstee, M. Ageing-in-place with the use of ambient intelligence technology: Perspectives of older users. Int. J. Med. Inf. 2011, 80, 310–331.
10. Demiris, G.; Hensel, B.K.; Skubic, M.; Rantz, M. Senior residents' perceived need of and preferences for "smart home" sensor technologies. Int. J. Technol. Assess. Health Care 2008, 24, 120–124
11. Korel, B.T.; Koo, S.G. Addressing context awareness techniques in body sensor networks. In Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW '07), Niagara Falls, ON, Canada, 21–23 May 2007; Volume 2, pp. 798–803.
12. Комаревич О.М. Виявлення аномалій у MicroGrid методами машинного навчання: Київ, 2018. 15-25с.
13. Gallant, S. I. (1990). Perceptron-based learning algorithms. IEEE Transactions on Neural Networks, vol. 1, no. 2, pp. 179–191.
14. Пискунов Н. С. Дифференциальное и интегральное исчисления для втузов, т. 2: Учебное пособие для втузов.—13-е изд.— М.: Наука, Главная редакция физико-математической литературы, 1985. — 560 с.
15. Tax, D. M. ., & Duin, R. P. . (1999). Support vector domain description. Pattern Recognition Letters, 20(11-13), 1191–1199. doi:10.1016/s0167-8655(99)00087-2 1192-1195c.
16. Support Vector Machines (SVM) Introductory Overview: http://www.statsoft.com/textbook/support-vector-machines
17. J. Mercer, Functions of positive and negative type and their connectionwith the theory of integral equations, Philos. Trans. Roy. Soc. London 1909